Stable Observable Behavior

Yuval Heller and Erik Mohlin

University of Oxford

19. March 2015

Stable Observable Behavior

Yuval Heller and Erik Mohlin*
Department of Economics, University of Oxford.

March 19, 2015

Abstract

We study stable behavior when players are randomly matched to play a game, and before the game begins each player may observe how his partner behaved in a few interactions in the past. We present a novel modeling approach and we show that strict Nash equilibria are always stable in such environments. We apply the model to study the Prisoner's Dilemma. We show that if players only observe past actions, then defection is the unique stable outcome. However, if players are able to observe past action profiles, then cooperation is also stable. Finally, we present extensions that study endogenous observation probabilities and the evolution of preferences.

JEL Classification: C72, C73, D01, D83. **Keywords**: Evolutionary stability, random matching, indirect reciprocity, secret handshake, submodularity, image scoring.

1 Introduction

In many economic situations people are involved in short-term interactions. The lack of future interactions between the agents limit the ability to directly punish partners who acted opportunistically, and incompleteness of contracts, non-verifiable information, court costs and other factors restrict the effectiveness of external enforcement. Agents in such interactions may obtain information about the partner's behavior in a few past interactions with other opponents, and may base their actions on this information. As a result, the behavior of agents may be influenced by the possibility of being observed by future partners and reputation considerations.

*Email: yuval.heller@economics.ox.ac.uk and erik.mohlin@economics.ox.ac.uk. We would like to express our deep gratitude to Vince Crawford, Eddie Dekel, Christoph Kuzmics, Bill Sandholm, Jörgen Weibull, and Peyton Young, and to seminar/workshop participants in Stockholm School of Economics, Bielefeld University, University of Pittsburgh, University of Oxford, NBER theory workshop in Wisconsin-Madison, for many useful comments.

A few examples for such interactions include booking a holiday rental, making a one-time order from a remote trader, hiring a nanny, and visiting a tourist attraction. Partial information about the past behavior of the partner is typically passed by word of mouth. Recently, such information is also conveyed by online sites that provide feedback from past interactions (e.g., eBay and Airbnb).

Model. Agents in a large population are randomly matched into pairs and play a symmetric one-shot game. Before playing the game, each agent may privately observe his partner's past behavior in k interactions. Specifically, each agent may either observe k past actions of his partner, or k past action-profiles of his partner and her past opponents, or a non-informative signal. The strategic behavior of an agent is described by a (stationary) *policy*: a mapping that assigns a mixed action to each possible message. A distribution of policies (called, *strategy*) describes the aggregate behavior: frequency of each faction in the population, and the policy that the agents of each faction follows.

Our preliminary result (Theorem 1) characterizes when every (stationary) strategy uniquely determines the distribution of action profiles played between each two groups in the populations (called, *outcome*), and when multiple outcomes might be consistent with a single strategy. We show that every strategy uniquely determines the outcome iff the expected number of actions that each agent observes about his partner is less than one.[1] [2] Therefore, a full description of a population state (when agents observe on average more than one action) is given by a *configuration*, a pair consisting of a strategy and a consistent outcome.

Solution Concept We present two static solution concepts that capture stability in a dynamic process of cultural learning (à la Maynard Smith & Price, 1973's notion of ESS). We imagine a large population in which each agent follows a policy. Occasionally a few agents receive an opportunity to change their policy. Generally such revisions go in the direction of the currently more successful policies (i.e. a payoff monotonic selection dynamics). However, with some small probability the revising agents may choose an arbitrary policy (and in this case they are referred to as *mutants*). We say that strategy is *evolutionary stable* if it satisfies three conditions: (1) it admits a unique consistent outcome (otherwise behavior may drift between the various consistent outcomes), (2) all policies in the support yield the same expected payoff (otherwise the more

[1] An example of multiple consistent outcomes is an environment in which agents play the Prisoner's Dilemma, each agent observes a single past action of his partner with probability one, and everyone plays the policy of mimicking the observed action. Any outcome (i.e., any frequency of cooperation) is consistent with this strategy.

[2] The intuition is that the strategy can be interpreted as a mapping from the outcome in the past to the outcome in the future, and that it is a contraction mapping iff agents observe (on average) less than one action.

successful polices will become more frequent), and (3) following an entry of a small number of agents who play a different strategy (called, *mutants*), the incumbents (strictly) outperform the mutants in any post-entry population state. The weaker notion, *weak stability*, which is used in our uniqueness result, allows multiple consistent outcomes, and only requires that the mutants will be outperformed in at least one consistent post-entry population state.

As is common in interactions with multiple stages, evolutionary (and weak) stability is too demanding in our setup: Typically not all feasible signals about the partner are observed on the equilibrium path, and as a result there exist equivalent strategies that differ from the incumbent strategy only after observing signals off the equilibrium path. This motivates us to follow Selten (1983) and to slightly weaken our definitions by only requiring stability in a converging sequence of perturbed games in which players rarely choose "wrong" actions (i.e., implementation errors similar to the trembles in extensive-form perfection of Selten, 1975) or "wrong" policies (i.e.., learning errors similar to the trembles in normal-form perfection); these trembles allow all signals to occur with positive probability and eliminate the issue of equivalent strategies. The adapted notions are called l*imit evolutionary stability* and *limit weak stability*.

Main Results Theorem 2 shows that any symmetric strict equilibrium of the underlying game is limit evolutionarily stable.[3] [4] The intuition is that a population in which everyone plays the strict equilibrium action a^* regardless of the observed signal is evolutionary stable: any mutant who does not play a^* with probability one against the incumbents is strictly outperformed if the mutants are sufficiently rare. Mutants who always play a^* against the incumbents, cannot identify other mutants (because the past behavior of mutants is the same as the incumbents), and thus must also play a^* among themselves.

We next consider *prisoner's dilemma games* in which each player decides simultaneously whether he cooperates or defects, where defecting yields a higher payoff for the defector and a lower payoff for the partner (see Table 2 in Section 6.1). A prisoner dilemma game is *submodular* (Takahashi, 2010) if the more likely a player is to cooperate, the better incentives his partner has to defect. Theorem 3 shows that always defecting is the unique limit weakly stable strategy in any (weakly) submodular prisoner's dilemma game if players observe past actions (rather than action profiles). The intuition is that in submodular games, it is always weakly better for a policy to defect against partners that are more likely to cooperate, but this implies that mutants who

[3] Moreover, the stability holds for *any* converging sequence of perturbations (with full support) in which players rarely make mistakes when choosing actions.

[4] In Section 5 we demonstrate that other refinements of Nash equilibrium are not necessarily stable. Specifically, we show that the unique symmetric Nash equilibrium in the Hawk-Dove game (which is stable without observability, and satisfies all the standard refinements of Nash equilibrium) is not stable for any positive level of observability.

always defect outperform the incumbents in any post-entry population state.

The next two results show how cooperation can be sustained as a stable outcome in the Prisoner's Dilemma game if either the players observe action-profiles (Theorem 4) or the underlying game is strictly supermodular (Theorem 5). In the former case we show that even if players observe a single action-profile, then there is a limit evolutionarily stable strategy that induces cooperation. In this strategy, players only defect if they observe that the partner was the sole defector, and cooperate otherwise. In the latter case we show that for strictly supermodular games even if the players observe a single action, then there is a limit evolutionarily stable strategy that induces cooperation. This strategy represents a polymorphic population in which one group in the population plays "tit-for-tat" (i.e. defects iff observing partner's past defect) and the other group always cooperate. In both cases, the stability relies on the fact that if a player observes his partner to defect in the past, then this partner is more likely to be a trembler who follows a policy that induces a higher defection probability.

Related Literature and Contribution Robson (1990) presented the *secret handshake* mechanism that can take a population from any inefficient state (say σ^*) to an efficient state (say σ'). According to this mechanism, a small group of mutants sends a special signal before interacting; the incumbents are assumed to ignore this signal and always play σ^*, while the mutants usually play σ^*, unless both players have sent the special signal, and in this case they both play σ'. Several papers applied this mechanism, and showed that with "cheap talk" with sufficiently rich language then stability implies efficiency (e.g., Wärneryd, 1991, Schlag, 1993, and Kim & Sobel, 1995). Theorem 2 shows that this mechanism is not effective when the observed signals are in fact past behavior: in this case non-efficient strict equilibria are stable. As argued in Section 7.4, this also casts some doubts on the use of secret handshake arguments to show that stability implies efficiency in the literature on the evolution of preferences (e.g., Dekel *et al.*, 2007).

In an influential paper, Nowak & Sigmund (1998), present the mechanism of *image scoring* to sustain cooperation in Prisoner Dilemma games through indirect reciprocity. In this mechanism each player observes several past actions of the partner, and he defects iff the partner's frequency of defection is above some threshold (see also the recent extension in Berger & Grüne, 2014). Our paper has two key contributions with respect to image scoring. First, Theorem 3 shows that image scoring cannot sustain stable cooperation in submodular prisoner's Dilemma games (this generalizes and formalizes existing criticism of image scoring - Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2003). Second, Theorem 5 presents a novel polymorphic variant of image scoring, and proves that it can sustain cooperation in supermodular prisoner's dilemma games.

In a seminal work Sugden (1986) presented the mechanism of *good standing* (see also its

extensions and applications in Kandori, 1992 and Ohtsuki & Iwasa, 2006). In this mechanism, all agents are initially in good standing; an agent looses his standing by defecting against a partner in good standing ; an individual lacking good standing can regain it by cooperating. In this mechanism players have to asses the standing of the partner, and this assessment is very demanding both informationally and cognitively, as it depends on long histories of play of many players (because the reputation of one agent depends on the reputations of his past partners). Theorem 4 presents a novel mechanism to support cooperation, in which each player only observes a single past action profile of his partner, without requiring any higher order information.

Finally, the closest paper to ours is Takahashi (2010) that studies the stability of cooperation when players observe the entire history of the partner's past play. Takahashi shows that: (I) there is a strict equilibrium that induces cooperation in the prisoner dilemma iff the game is supermodular, and (II) there is a sequential equilibrium that induces cooperation in any prisoner dilemma game. Our key contribution with respect to Takahashi (2010) is three-fold. First, we develop a novel methodology that allows to study environments in which players have partial information about the partner's past play. Second, we show that cooperation can be sustained as a stable outcome in supermodular games also when players observe a single past action of the partner (rather than requiring the observation of the full history of play). Finally, we show that only defection is stable in submodular games (and, in particular, the sequential equilibria in Takahashi, 2010 are unstable).

Variants and Extensions. We present four variants and extensions (Section 7). The first assumes that both signals are observed by both players. This assumption may fit better trade in on-line sites in which the feedback on the past behavior of the agents is public. Minor adaptations to the proofs show that all the results of the main model hold also in this setup. In addition, we show (Theorem 6) that public signals allow to support cooperation as a stable outcome also if players only make mistakes when choosing actions (rather than when choosing policies).

Next we discuss how to extend our results to a setup in which agents may choose non-stationary policies that may condition their play on their own past behavior. The third extension endogenizes the observation probability by assuming that each player chooses an effort level that determines the probability to observe his partner's past.

The last variant enriches the model to study the evolution of preferences. Specifically, we assume that each agent is endowed with subjective preferences that may differ from the objective fitness. A population state is now a triple: a distribution of subjective preferences, a policy for each preference (which maximizes the agent's subjective utility given his information about the partner), and a consistent outcome. We extend our first result to this setup, and show that strict

equilibria (including non-efficient ones) are always stable in this setup. This contradicts the main stylized result in the existing literature on the evolution of preference that only efficient outcomes can be stable if there is sufficient observability (where, unlike our model, agents directly observe the partner's matrix payoff).

Structure. Section 2 presents the model. In section 3 we define population states. Section 4 presents our solution concepts. In Section 5 we present our first result (any strict equilibrium is stable). Section 6 focuses on the Prisoner's Dilemma and shows the remaining main results: only defection is stable when observing actions, while cooperation is also stable when observing action profiles or with supermodularity. We present variants and extensions in Section 7, and we conclude in Section 8. The formal proofs appear in the appendix.

2 Model

2.1 Underlying Game

We present a reduced form static analysis of a dynamic evolutionary process of cultural learning in a large population of agents.[5] The agents in the population are randomly matched into pairs and play a symmetric one-shot game G. Formally, let $G = (A, \pi)$ be a two-player symmetric normal-form game, where A is a finite set of actions ($|A| \geq 2$), and $\pi : A \times A \to \mathbb{R}$ is the payoff function. As is standard in the evolutionary game theory literature, we interpret the payoffs as representing "success" or "fitness".

Let $\Delta(A)$ denote the set of mixed actions (distributions over A), and let π be extended to mixed actions in the usual way. We use the letter a (α) to denote a typical pure (mixed) action. With slight abuse of notation let $a \in A$ also denote the element in $\Delta(A)$, which assigns probability 1 to a. We adopt this convention for all probability distributions throughout the paper.

Remark 1. The assumption that the underlying game G is symmetric is essentially without loss of generality (if the game is played within a single population). As is standard in the evolutionary literature (e.g., Selten, 1980), asymmetric contests can be symmetrized by looking at a symmetric game in which an agent chooses an action at the ex ante stage (i.e., before being assigned to one of the roles in the game).

[5] Alternatively, the formal model may be interpreted in terms of a biological evolutionary process in which the behavior of the agents is influenced by genetic inheritance.

2.2 Observation Structure and Environment

We study an environment in which agents are randomly matched in each turn to play game G. Before playing the game, each player may observe a few past actions or action profiles played by his partner. Specifically, each player privately observes:

1. With probability $p_1 \in [0,1]$: $k \geq 1$ random actions played in the past by his partner.
2. With probability $p_2 \in [0, 1-p_1]$: $k \geq 1$ random action profiles played in past interactions of his partner and her opponents.
3. With the remaining probability $p_0 \equiv 1 - p_1 - p_2$: a non-informative signal ϕ.

We refer to the tuple of parameters $\omega = (k, p_1, p_2)$ as the *observation structure*. An *environment* is a pair $E = (G, \omega)$, where G is a (two-player symmetric) underlying game, and ω is the observability structure. Let M denote the set of possible *signals* (or *messages*) about the partner's past: $M = A^k \cup (A \times A)^k \cup \phi$. Let m denote a typical message (i.e., an element of M).

Remark 2. The assumptions that (1) the number of observations is fixed (either 0 or k), and (2) agents observe either only actions or only action profiles, are made only to simplify the notation. All of the results can be extended to a setup in which the number of observations has a larger support, and agents observe a combination of actions and action profiles..

2.3 Strategies

A *policy* is a mapping $s : M \to \triangle(A)$ that assigns a mixed action to each possible message. We interpret $s_m(a) = s(m)(a)$ as the probability that a player who follows policy s plays action a after observing message m. We also let a denote the policy $s \equiv a$ that plays action a regardless of the signal.

Let S denote the set of all policies, and let $\Sigma \equiv \triangle(S)$ denote the set of finite support distributions over the set of policies. An element $\sigma \in \Sigma$ is called a *strategy* (or *policy distribution*). Let $\sigma(s)$ denote the probability that strategy σ assigns to policy s. Given a strategy $\sigma \in \Sigma$, let $C(\sigma)$ denote its support (i.e., the set of policies such that $\sigma(s) > 0$). We interpret a strategy $\sigma \in \Sigma$ as representing a population in which $|C(\sigma)|$ policies coexist, each agent is endowed with one of these policies according to the distribution of σ. When $|C(\sigma)| = 1$, we identify the strategy with the unique policy in its support (i.e., $\sigma \equiv s$), in line with the convention adopted above.

Remark 3. Note that our model focuses on stationary policies in which the behavior of an agent depends only on the signal about the partner, but agents cannot condition their behavior on their own past play or on time. We discuss how to relax this assumption in Section 7.2.

3 Population States

In this section we present the static notion of configuration, which we use to fully describe the state of the population.

3.1 Outcomes

Given a finite set of policies $\tilde{S} \subset S$, an *outcome* $\eta : \tilde{S} \times \tilde{S} \to \triangle(A)$ is a mapping that assigns to each pair of policies $s, s' \in \tilde{S}$ a mixed action $\eta_s(s')$, which is interpreted as the mixed action played by an agent with policy s conditional on being matched with a partner with policy s'. Let $O_{\tilde{S}} \equiv (\triangle(A))^{\left(\tilde{S}\times\tilde{S}\right)}$ denote the set of all outcomes defined over the set of policies $\tilde{S}$. The strategy and the outcome together determine the payoffs earned by each agent in the population.

Outcome $\eta \in O_{\tilde{S}}$ is *pure* if there exists action $a \in A$ such that $\eta_s(s') = a$ for each $s, s' \in \tilde{S}$. We denote such a pure outcome as $\eta \equiv a$.

We present a few definitions for a given strategy $\sigma \in \Sigma$, an outcome $\eta \in O_{C(\sigma)}$, and a policy $s \in C(\sigma)$. Let $\eta_{s,\sigma} \in \triangle(A)$ be the mixed action played by an agent with policy s when being matched with a random partner sampled from σ. Formally, for each action $a \in A$:

$$\eta_{s,\sigma}(a) = \sum_{s' \in C(\sigma)} \sigma(s') \cdot \eta_s(s')(a).$$

Let $\psi_{s,\sigma,\eta} \in \triangle(A \times A)$ be the (possibly correlated) mixed action profile that is played when an agent with policy s is matched with a random opponent sampled from σ. Formally, for each action profile $(a, a') \in A \times A$, where a is interpreted as the action of the agent with policy s, and a' is interpreted as the action of his partner:

$$\psi_{s,\sigma,\eta}(a, a') = \sum_{s' \in C(\sigma)} \sigma(s') \cdot \eta_s(s')(a) \cdot \eta_{s'}(s)(a').$$

3.2 Consistent Outcomes

Fix an observation structure $\omega = (k, p_1, p_2)$. Suppose that the population is currently described by the policy distribution σ and the outcome η. When individuals are drawn to play the game their actions are determined by their policy and their observed signal. These action profiles induce a new outcome $f_\sigma(\eta)$. We will require outcomes to be consistent in the sense that they generate observations that induces the current outcome to persist. Formally, given strategy $\sigma \in \Sigma$, let

$f_\sigma : O_{C(\sigma)} \to O_{C(\sigma)}$ be the mapping between outcomes that is induced by σ.

$$\begin{aligned}(f_\sigma(\eta))_s(s')(a) &= p_0 \cdot s(\phi)(a) + p_1 \cdot \sum_{(a_i)_{i\le k}\in A^k} \prod_{i\le k} \eta_{s',\sigma}(a_i) \cdot s\left((a_i)_{i\le k}\right)(a) \\ &\quad + p_2 \cdot \sum_{(a_i,a_i')_{i\le k}\in A\times A} \prod_{i\le k} \psi_{s',\sigma,\eta}(a_i,a_i') \cdot s\left((a_i,a_i')_{i\le k}\right)(a).\end{aligned}$$

An outcome $\eta \in O_{C(\sigma)}$ is consistent with strategy σ if it is a fixed point of this mapping: $f_\sigma(\eta) \equiv \eta$. The following standard lemma shows that each strategy admits consistent outcomes.

Lemma 1. *For each strategy $\sigma \in \Sigma$ there exists a consistent outcome η.*

Proof. Observe that the space $O_{C(\sigma)}$ is a convex and compact subset of a Euclidean space, and that the mapping $f_\sigma : O_{C(\sigma)} \to O_{C(\sigma)}$ is continuous. Brouwer's fixed-point theorem implies that the mapping σ has a fixed point $\eta*$, which is a consistent outcome by definition. □

The following simple example shows that a strategy may admit multiple consistent outcomes.

Example 1. Assume that the underlying game has two actions $A = \{c, d\}$, and that each player observes a single past action of his opponent with probability one (i.e., $k = 1$, $p_1 = 1$, $p_2 = 0$). Let $\tilde{s}$ be the following "tit-for-tat" policy: $\tilde{s}(a) = a$ for each $a \in A$ (i.e., an individual that follows this policy plays the observed past action of his opponent). Note that any outcome $\eta \in O_{\tilde{s}}$ is consistent with the strategy that assigns mass 1 to policy $\tilde{s}$. Note that if $p_1 < 1$ (while keeping $p_2 = 0$ and $k = 1$) then it is relatively simple to show that the unique consistent outcome is the distribution played after observing a non-informative signal: $\eta_s(s)(c) = \tilde{s}(\phi)(c)$.

3.3 Uniqueness of Consistent Outcomes

Proposition 1 characterizes which observability structures induce unique consistent outcomes; that is, structures in which every strategy admits a unique consistent outcome. It turns out that an observability structure induces unique consistent outcomes iff the expected number of actions that each agent observes about his partner in each round $(k \cdot p_1 + 2 \cdot k \cdot p_2)$ is at most one.

Theorem 1. *The following conditions are equivalent: (1) Every strategy in environment $E = ((A,\pi),(k,p_1,p_2))$ admits a unique consistent outcome; and (2) $k \cdot p_1 + 2 \cdot k \cdot p_2 \le 1$ and $p_1 < 1$.*

Proof. We begin by proving that "¬2" implies "¬1" (which is logically equivalent to proving that "1" implies "2"). Assume that $k \cdot p_1 + 2 \cdot k \cdot p_2 > 1$.[6] Let $a \neq a'$ be different actions. Let s^* be

[6]Recall that the case of $k = p_1 = 1$ was dealt with in Example 1 with multiple consistent outcomes.

the following policy: play a if the observed actions include a, and play a' otherwise. Consider the policy distribution in which all agents follow policy s^*. Consider the outcome η_x that assigns probability $0 \leq x \leq 1$ to action a and the remaining probability $(1-x)$ to action a'. Note that outcome η_x is consistent with s^* iff

$$x = Pr\left(\text{observing } a\right) = p_1 \cdot \left(1 - (1-x)^k\right) + p_2 \cdot \left(1 - (1-x)^{2 \cdot k}\right).$$

It is immediate that $x = 0$ always solves this equation, and thus η_0 is a consistent outcome. Next, note that when $x > 0$ is close to 0 the RHS can be (Taylor-)approximated by:

$$\left(p_1 \cdot k + p_2 \cdot (2 \cdot k)\right) \cdot x > 1 \cdot x.$$

For $x = 1$ the RHS is $p_1 + p_2 \leq 1$, so if $p_1 + p_2 = 1$ then $x = 1$ is also a solution and if $p_1 + p_2 < 1$ then by continuity of the RHS, there is some $x \in (0,1)$ that solves the equation. Thus there is $\eta_x \neq \eta_0$ that is also a consistent outcome of s^*.

In what follows, we sketch the proof of the opposite direction: "2" implies "1", while leaving the formal details to Appendix A.1. Let σ be an arbitrary strategy, and let η and η' be two outcomes. Let $\rho_M\left(\sigma, \eta\right) \in \Delta\left(M\right)$ denote the distribution of signals (about the past actions of a random partner) induced by strategy σ and outcome η. Observe that the distance (see Appendix A.1 for the formal definition of the norm) between the outcomes $f_\sigma\left(\eta\right)$ and $f_\sigma\left(\eta'\right)$ is bounded by the distance between the distributions of signals:

$$\left\| f_\sigma\left(\eta\right) - f_\sigma\left(\eta'\right) \right\| \leq \left\| \rho_M\left(\sigma, \eta\right) - \rho_M\left(\sigma, \eta'\right) \right\|.$$

This is because the mapping f_σ can induce different outcomes only to the extent that the observed signals were different. Next observe that the distance between the two signal distributions is bounded by the distance between the two outcomes times the expected number of messages.

$$\left\| \rho_M\left(\sigma, \eta\right) - \rho_M\left(\sigma, \eta'\right) \right\| \leq \left(k \cdot p_1 + 2 \cdot k \cdot p_2\right) \cdot \left\| \eta - \eta' \right\|.$$

This is because two random observed actions are different in probability $\|\eta - \eta'\|$, and the probability of k observed actions to differ in at least one action is at most k times $\|\eta - \eta'\|$ (with strict inequality if $k > 1$). These two inequalities imply that if $k \cdot p_1 + 2 \cdot k \cdot p_2 < 1$ (or if $k \cdot p_1 + 2 \cdot k \cdot p_2 = 1 < k$) then f_σ is a contraction mapping (implying unique consistent outcome): $\left\| f_\sigma\left(\eta\right) - f_\sigma\left(\eta'\right) \right\| < \left\| \eta - \eta' \right\|$. □

3.4 Configurations and Payoffs

Proposition 1 shows that a strategy fully describes a population state if and only if agents observe on average less than one action. As we are interested in studying also the case in which agents observe more actions, we use the notion of configuration, a pair consisting of a strategy and a consistent outcome, to describe the state of the population. Formally:

Definition 1. A *configuration* (or *population state*) is a pair (σ, η), where $\sigma \in \Sigma$ is a strategy and $\eta \in O_{C(\sigma)}$ is a consistent outcome (i.e., $f_\sigma(\eta) \equiv \eta$).

Given a configuration (σ, η) and a policy $s \in C(\sigma)$, let $\pi_s(\sigma, \eta)$ be the payoff of a player who follows policy s in population state (σ, η):

$$\pi_s(\sigma, \eta) = \sum_{(a,a') \in A \times A} \pi(a, a') \cdot \mu_{s,\sigma,\eta}(a, a').$$

Given a strategy σ' with a weakly smaller support than σ ($C(\sigma') \subseteq C(\sigma)$), let $\pi_{\sigma'}(\sigma, \eta)$ be he payoff of a player with a policy sampled according to σ' in population state (σ, η):

$$\pi_{\sigma'}(\sigma, \eta) = \sum_{s' \in C(\sigma')} \sigma'(s') \cdot \pi_{s'}(\sigma, \eta).$$

We say that configuration (σ, η) is *balanced* if $\pi_s(\sigma, \eta) = \pi_{s'}(\sigma, \eta)$ for every $s, s' \in C(\sigma)$. In that case, we write the uniform payoff as $\pi(\sigma, \eta)$.

4 Solution Concepts

In this section we adapt and extend the notions of evolutionary stability (Maynard-Smith, 1974) and limit evolutionary stability (Selten, 1983; Heller, 2014) to deal with the environments with observable past actions. We present two main static solution concepts: a strong notion to be used in existence results, and a weak notion to be used in uniqueness results.

4.1 Post-Entry Populations

Our static concepts are intended to capture stable behavior in a dynamic process of cultural learning. We imagine a large population of agents. At each point in time every agent in the population has a policy that he currently follows. Occasionally a few agents are drawn to receive the opportunity to change their policy. Generally such revisions go in the direction of the currently more successful policies (i.e. payoff monotonic selection dynamics). However, with some small

probability the revising agents may choose an arbitrary policy, which may not have been present in the population before. Such deviations from local optimality may be due to mistakes or to conscious experimentation. We assume that these deviations are rare and involve only a small fraction of the population, which we will refer to as *mutants*.

Our stability notions consider incumbents who follow strategy σ^* until a small group of mutants (with small mass $\epsilon > 0$), who play a different strategy σ', enter the population . Following the entry, the distribution of policies in the population is a mixture strategy that gives weight of $1-\epsilon$ to the incumbents' strategy and weight of ϵ to the mutants' strategy. The behavior of the population following such an entry is assumed to converge to a consistent outcome of this mixture strategy.

Formally, Given $0 < \epsilon < 1$ and two strategies $\sigma^*, \sigma' \in \Sigma$ with relative masses of $1-\epsilon$ and ϵ, let $\sigma_\epsilon = \sigma_{\sigma^*,\epsilon,\sigma'}$ denote the *mixture strategy*:

$$\sigma_\epsilon(s) = (1-\epsilon)\cdot\sigma^*(s) + \epsilon\cdot\sigma'(s) \quad \text{for each } s \in C(\sigma)\cup C(\sigma'),$$

and let a *post-entry configuration* be any configuration consisting of the mixture strategy and a consistent outcome: $(\sigma_\epsilon, \eta_\epsilon)$.

4.2 Evolutionary Stability

In this subsection we define a strong notion of evolutionary stability, which will be used in the existence results in the paper (after adapting it to small perturbations in Section 4.4).

A strategy is evolutionary stable if (1) it admits a unique consistent outcome, (2) all policies yield the same payoff, and (3) and small group of mutants is outperformed in any post-entry population. Formally:

Definition 2. The number $\bar{\epsilon} > 0$ is an invasion barrier for strategy σ^* if for each strategy $\sigma' \neq \sigma^*$, each $\epsilon \in (0, \bar{\epsilon})$ and each post-entry configuration $(\sigma_\epsilon = (1-\epsilon)\cdot\sigma^* + \epsilon\cdot\sigma', \eta_\epsilon)$ the following inequality holds:

$$\pi_{\sigma'}(\sigma_\epsilon, \eta_\epsilon) < \pi_{\sigma^*}(\sigma_\epsilon, \eta_\epsilon).$$

σ^* has a *uniform invasion barrier* if there is $\bar{\epsilon} > 0$ such that $\bar{\epsilon}$ is an invasion barrier for σ^*.

Definition 3. Strategy σ^* is *evolutionarily stable* if it satisfies the following conditions:

1. Strategy σ^* admits a unique consistent outcome η^*.

2. The configuration (σ^*, η^*) is balanced.

3. Strategy σ^* has a uniform invasion barrier.

The motivation for the first condition (unique consistent outcome) is that otherwise the behavior is inherently unstable as it may drift between the various consistent outcomes. The second condition is necessary for internal stability: if the policy distribution is not balanced, then the frequency of the agents who follow the more successful policy will increase.

The third condition is analogous to Maynard Smith & Price (1973)'s notion of evolutionary stability in standard games (without observability).[7] The motivation for this condition is that an evolutionarily stable strategy, if adopted by the population, cannot be invaded by any alternative strategy that is initially rare. It is well known that evolutionarily stable strategies are asymptotically stable (a population starting nearby eventually converges to it) in a large variety of smooth payoff-monotonic selection dynamics that capture many plausible processes of cultural learning (see, e.g., Taylor & Jonker (1978); Cressman (1997); Sandholm (2010)). Moreover, the converse is also true in the sense that only evolutionary stable strategies are asymptotically stable in all variants of the replicator dynamics regardless of which randomizations are allowed at the individual level (Cressman, 1990).

4.3 Weak Stability

In this subsection we define a weak notion of evolutionary stability, which will be used in the uniqueness results in the paper.

A balanced configuration is weakly stable if any sufficiently small group of mutants is outperformed in at least one post-entry configuration. Formally:

Definition 4. Number $\bar{\epsilon} > 0$ is a *weak invasion barrier* for strategy σ^*if for each $\epsilon \in (0, \bar{\epsilon})$, and each mutant strategy $\sigma' \in \Sigma$, there exists a post-entry configuration $(\sigma_\epsilon, \eta_\epsilon)$ such that:

$$\pi_{\sigma'}(\sigma_\epsilon, \eta_\epsilon) < \pi_{\sigma^*}(\sigma_\epsilon, \eta_\epsilon).$$

Strategy σ^* has a *weak invasion barrier* if there is $\bar{\epsilon} > 0$ that is a weak invasion barrier for σ^*.

Definition 5. Configuration (σ^*, η^*) is *weakly stable* if it is balanced and strategy σ^* has a weak invasion barrier.

Definition 5 is weaker than Definition 3 in two ways:

1. Strategy σ^* may admit multiple consistent outcomes.

[7] We follow the reformulation of evolutionary stability with uniform invasion barrier a la Hofbauer *et al.* (1979), see also (Weibull, 1995, Proposition 2.5).

2. Following an entry of mutants, there might be several post-entry configurations, and we only require the mutants to be outperformed in one of these configurations.

It is immediate that any evolutionary stable strategy (together with its unique outcome) is a weakly stable configuration, and that both definitions coincide when we have unique consistent outcomes (i.e., $k \cdot (p_1 + 2 \cdot p_2) < 1$). Moreover, when there is no observability (i.e., $p_1 = p_2 = 0$) the definitions coincide with Maynard-Smith and Price (1973)'s notion of evolutionary stability.

4.4 Limit Evolutionary Stability

Each interaction in our environment includes two stages: (1) observing signals about the partner, and (2) playing an action in the underlying game. As is common in interactions with multiple stages, evolutionary stability is too demanding in our setup: Typically not all signals about the partner are observed on the equilibrium path, and as a result there exist equivalent strategies that induce the same behavior as the incumbent strategy and differ only following signals that are off the equilibrium path. This motivates us to slightly weaken the notion of evolutionary stability to require evolutionary stability only in a converging sequence of perturbed games in which players rarely "tremble" and choose the "wrong" policy or the "wrong" action (a la Selten, 1983's notion of limit ESS as reformulated in Heller, 2014).

A perturbation is a mapping that assigns minimal inevitable probabilities to choose some of the actions and some of the policies. Formally:

Definition 6. A *perturbation is a tuple* $\zeta = (\xi, \mathcal{S}, \lambda)$ where:

1. $\xi : A \to \mathbb{R}^+$ is a function that assigns a non-negative number for each action such that $\sum_{a \in A} \xi(a) < 1$. Let $S_\xi \subseteq S$ the set of policies that assign probability of at least $\xi(a)$ to each action a after any observed signal.

2. $\mathcal{S} \subseteq S_\xi$ is a finite set of policies, and $\lambda : \mathcal{S} \to \mathbb{R}^+$ is a function that assigns a non-negative number for each policy in $\mathcal{S}$ such that $\sum_{s \in \mathcal{S}} \lambda(s) < 1$.

Let $E(\zeta)$ denote the *perturbed environment* that results from perturbing an environment E by perturbation ζ. That is, the environment $E(\zeta)$ is the same as E except that the set of feasible strategies is limited to those that assign the minimal probabilities induced by ζ. Formally, let the convex set of *feasible strategies* $\Sigma(\zeta) \subseteq \Sigma$ be defined as follows. Strategy $\sigma \in \Sigma$ is included in $\Sigma(\zeta)$ iff: (1) $C(\sigma) \subseteq S_\xi$, and (2) For each $s \in \mathcal{S}$, $\sigma(s) \geq \lambda(s)$. That is, players tremble when choosing actions according to ξ, and tremble when learning their policies according to $(\mathcal{S}, \lambda)$. Let $L(\zeta)$ denote the *maximal tremble* of $\zeta = (\xi, \mathcal{S}, \lambda)$: $L(\zeta) = \max(\max_{a \in A} \xi(a), \max_{s \in \mathcal{S}} \lambda(s))$.

Remark 4. Note that trembles when choosing the policy are closely related to the trembles used in the notion of normal-form perfection (Selten, 1975), while trembles when choosing actions are equivalent to the more "common" trembles used in the definitions of extensive-form perfection and limit ESS (Selten, 1983). See Selten (1975, Section 13) and van Damme (1987, Section 6.4) for discussing the differences between normal-form perfection in which players tremble when choosing a strategy for the entire game (a policy in our setup), and extensive-form perfection in which players tremble when choosing actions at information sets.

A limit evolutionarily stable strategy is the limit of evolutionarily stable strategies in a converging sequence of perturbed games. Formally:

Definition 7. Sequence of configurations (σ_n, η_n) converge to configuration (σ^*, η^*) if:

1. For each policy $s \in S$: $\sigma_n(s) \to \sigma^*(s)$.
2. For each two policies $s, s' \in C(\sigma^*)$: $(\eta_n)_s(s') \to \eta^*_s(s')$.

Definition 8. Strategy σ^* is *limit* evolutionarily stable if:

1. σ^* admits a unique consistent outcome η^*, and the configuration (σ^*, η^*) is balanced.
2. There exists sequence $(\zeta_n, (\sigma_n, \eta_n))_{n\in\mathbb{N}}$ such that:
   (a) ζ_n is a sequence of perturbations converging to the unperturbed game: $L(\zeta_n) \to 0$.
   (b) Each strategy σ_n is feasible in $E(\zeta_n)$ (i.e., $\sigma_n \in \Sigma(\zeta_n)$), each pair (σ_n, η_n) is a configuration, and the sequence of configurations (σ_n, η_n) converges to (σ^*, η^*).
   (c) There exist $\bar{\epsilon} > 0$ and $n_0 \in \mathbb{N}$ such that $\bar{\epsilon}$ is an invasion barrier for strategy σ_n (in environment $E(\zeta_n)$) for each $n \geq n_0$.

Finally, limit weakly stable configuration is defined as the limit of weakly stable configurations in a converging sequence of perturbed games. Formally:

Definition 9. Configuration (σ^*, η^*) is *limit weakly stable* if it is balanced, and there exists $\bar{\epsilon} > 0$ and sequence $(\zeta_n, (\sigma_n, \eta_n))_{n\in\mathbb{N}}$ such that:

1. ζ_n is a sequence of perturbations converging to the unperturbed game: $L(\zeta_n) \to 0$.
2. Each strategy σ_n is feasible in the perturbed game $\Gamma(\zeta_n)$: $\sigma_n \in \Sigma(\zeta_n)$, each pair (σ_n, η_n) is a configuration, and the sequence of configurations (σ_n, η_n) converges to (σ^*, η^*).
3. There exist $\bar{\epsilon} > 0$ and $n_0 \in \mathbb{N}$ such that $\bar{\epsilon}$ is a weak invasion barrier for strategy σ_n (in environment $E(\zeta_n)$) for each $n \geq n_0$.

Observe that both definitions of limit evolutionary stability and weak evolutionary stability coincide with each other and with Heller (2014)'s notion of uniform limit ESS.[8]

4.5 Action-Strict Evolutionary Stability

The choice of perturbations is critical for whether it will be possible to construct post-entry configurations that are close and at which mutants are outperformed. The above definition only requires that there is one sequence of perturbations with the desired properties. However, in some environments we may seek a more robust definition that requires stability with respect to a whole class of perturbations. In some contexts it is reasonable to assume that trembles are more likely to occur when choosing actions than when revising and choosing policies. In particular, this is the implicit assumption in the commonly used refinements of (extensive-form) perfect equilibrium an sequential equilibrium. We present a more robust definition for such contexts that requires stability with respect to all perturbations in which players tremble when choosing actions, and these trembles have full support. The definition is closely related to the notion of strict limit ESS (Heller, 2015) and to strict perfection (Okada, 1981). Formally:

Definition 10. A *perturbation* $\zeta = (\xi, \mathcal{S}, \lambda)$ is an *action perturbation with full support* if:

1. Trembles when choosing actions have full support : $\xi(a) > 0$ for each action $a \in A$.
2. There are no trembles when choosing policies: $\lambda \equiv 0$.

We identify an action perturbation with full support with its first component ξ.

Definition 11. Strategy σ^* is *action-strict evolutionarily stable* if:

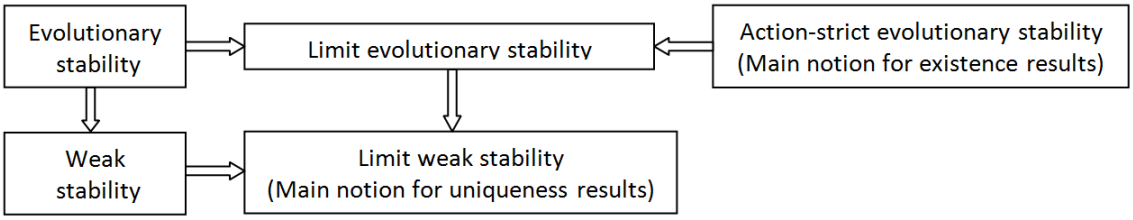
1. σ^* admits a unique consistent outcome η^*, and the configuration (σ^*, η^*) is balanced.
2. There exists $\bar{\epsilon} > 0$ such that for each converging sequence of action perturbations with full support $(\xi_n)_{n\in\mathbb{N}}$ (i.e., $L(\xi) \to 0$), there exists a sequence $(\sigma_n, \eta_n)_{n\in\mathbb{N}}$ satisfying:
   (a) Each strategy σ_n is feasible in the perturbed environment $E(\zeta_n)$: $\sigma_n \in \Sigma(\zeta_n)$, each pair (σ_n, η_n) is a configuration, and the sequence of configurations (σ_n, η_n) converges to (σ^*, η^*).
   (b) There exists $n_0 \in \mathbb{N}$ such that $\bar{\epsilon}$ is an invasion barrier for strategy σ_n in environment $E(\zeta_n)$ for each $n \geq n_0$.

[8] See, Heller (2014) for a discussion why there should be a uniform invasion barrier for all the σ_n-s, rather than allowing invasion barriers that depend on n (as is the case in the original notion of limit ESS in Selten, 1983).

4.6 Summary

We conclude this section by presenting in Figure 1 the logical relations between the different notions (which are immediate). We will use the notion of action-strict evolutionary stability for most of the existence results, and the notion of limit weak stability for the uniqueness results.

Figure 1: The Logical Relations Between the Different Solution Concepts.



5 Stability of Strict Equilibria

In this section we show that each strict Nash equilibrium (a^*, a^*) of the underlying game is stable in a strong sense given any observation structure. Specifically, the strategy a^* (i.e., the strategy that assigns mass one to the policy that always plays a^* regardless of the observed signal) is an action-strict evolutionarily stable strategy.

The intuition is as follows. Consider a slightly perturbed environment in which players sometimes tremble when choosing actions, and all signals are observed with positive probability. Consider a configuration in which all the (non-trembling) incumbents play the equilibrium action a^* regardless of the signal about the partner's past. Any mutant strategy must play a different action with positive probability against the incumbents, and thus it yields a lower payoff when being matched against the incumbents. If the mutants are sufficiently rare, then this loss cannot be compensated by a gain when facing other mutants. This implies that the configuration in which everyone always plays a^* is limit evolutionarily stable. Formally (proof is in the appendix):

Definition 12. The action a^* is a strict (symmetric) Nash equilibrium of the underlying game $G = (A, \pi)$ if for each action $a \neq a^*$: $\pi(a^*, a^*) > \pi(a, a^*)$.

Theorem 2. *If action a^* is a strict equilibrium then the strategy a^* is an action-strict evolutionarily stable strategy.*

An interesting question is whether Theorem 2 can be strengthened to weaker solution concepts (of the underlying game) than strict equilibrium. The following example shows that this is not the case: the unique symmetric Nash equilibrium of the underlying game is not stable in

any environment with observability ($p_0 < 1$). Note that a unique symmetric equilibrium of a symmetric game satisfies all the standard (non-evolutionary) refinements of Nash equilibrium.

Table 1: An Example of a Hawk-Dove (Chicken) Game

	d	h
d	1, 1	0.5, 1.5
h	1.5, 0.5	0, 0

Example 2 (Instability of a unique symmetric Nash equilibrium)**.** Consider the game of Hawk-Dove (or Chicken) presented in Table 1 in which the players have two actions: d ("dove") and h (hawk). Each action is the strict best-reply to the other action, and $\alpha^* = (0.5, 0.5)$[9] is the unique symmetric Nash equilibrium, as well as the unique evolutionarily stable strategy of the underlying game.[10] We now show why strategy α^* with its unique consistent outcome , $\eta^* \equiv \alpha^*$ is not limit weakly stable if $p_0 < 1$. To simplify the argument we assume that each agent may only observe a single action (i.e., $k = 1$, $p_1 > 0 = p_2$), but the argument can be be extended to observation of several actions and action profiles (and to any Hawk-dove game). Consider a mutant strategy that assigns equal weights to three policies: (1) always play h, (2) always play d, and (3) play the action is that the opposite of the observed message, and play each action with equal probability if ϕ was observed. These mutants obtain the same payoff as incumbents when facing incumbents (because all actions yield the same payoff against α^*), but a strictly higher payoff relative to the incumbents when facing other mutants (because when two mutants are matched they play the inefficient action profile (h, h) with probability of only $\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \cdot \frac{1}{4} < \frac{1}{4}$, while when an incumbent and a mutant are matched they play (h, h) with probability $\frac{1}{4}$). This implies that the mutants outperform the incumbents in any post-entry population (and the argument works also for sufficiently close perturbed environments).

6 Prisoner's Dilemma

In this section we study Prisoner's Dilemma games. Theorem 3 shows that defection is the unique stable outcome in submodular Prisoner's Dilemma game when players only observe actions.

[9] The vector $(\alpha_1, ..., \alpha_n)$ denotes the mixed action that assigns probability α_i to the i-th action.

[10] Recall that asymmetric equilibria cannot be played in our setup in which the agents cannot condition their play on being the row/column player.

The remaining results show how to support stable cooperation when observing action profiles (Theorem 4) or when the underlying game is supermodular (Theorem 5).

6.1 Payoff Matrix and Submodularity

Table 2: Matrix Payoff of Prisoner's Dilemma Game ($g, l > 0$)

	c	d
c	1 1	$1+g$ $-l$
d	$-l$ $1+g$	0 0

Table 2 presents the matrix payoff of a Prisoner's Dilemma game that depends on two positive parameters g and l. When both players play action c (*cooperate*) they both get a high payoff (normalized to one), and when they both play action d (*defect*) they get a low payoff (normalized to zero). When a single player defects he obtains a payoff of $1+g$ (i.e., an additional payoff of g) while his opponent gets $-l$. Note that d is strictly dominant and (d, d) is the unique Nash equilibrium (and a strict equilibrium) of the game.

Following Takahashi (2010), we say that a Prisoner's Dilemma game is (weakly) *submodular* if $g \geq l$: the more likely a player is to cooperate, the better incentives his partner has to defect. In the opposite case ($g < l$) we say that the game is *strictly supermodular*: the more likely a player is to cooperate, the better incentives his partner has to cooperate.[11]

6.2 Only Defection is Stable (Observing Actions and Submodularity)

This section deals with the submodular Prisoner's Dilemma games and observations of past actions (rather than action profiles). Under these assumptions we obtain a sharp uniqueness result: always defecting is the unique weakly stable strategy.[12] Formally:

Theorem 3. *Assume that G is a submodular Prisoner's Dilemma game (i.e. $g \geq l$), and that $p_2 = 0$. If (σ, η) is a limit weak evolutionarily stable configuration, then $\eta \equiv d$.*

Sketch of Proof (see formal proof in Appendix A.3). Assume that (σ, η) is a limit weakly stable configuration. The payoff of a policy in the Prisoner Dilemma can be divided into two components:

[11] Dixit (2003) called $g \geq l$ the *offensive* case and $g < l$ the *defensive* case.

[12] Recall, that Theorem 2 shows that always defecting is is an action-strict evolutionary stable, which implies limit weak stability.

(1) direct component - defecting yields additional g (l) points if the partner cooperates (defects); (2) indirect component - the policy's average probability of defection determines the distribution of actions observed by the partners, and through this, influences the partner's probability to defect. For each fixed average probability of defection q the submodularity of the game implies that the optimal policy among all the policy that defect with average probability q is the one that defects with higher probability against partners that are more likely to cooperate. □

This implies that there is some optimal defection probability $0 \leq \hat{q} \leq 1$, such that the policy $s_{\hat{q}}$ that defects with average probability $\hat{q}$ and is more likely to defect against partners that are more likely to cooperate outperforms any other policy. This implies that (σ, η) can be limit weakly stable configuration only if all non-trembling follow policy $s_{\hat{q}}$. If $\hat{q} < 1$, then mutants who always defect would outperform the incumbents: they would achieve at least the same payoff (relative to the incumbents) when being matched against an incumbent (as their own tendency to defect would induce a partner following $s_{\hat{q}}$ to cooperate against them) and a higher payoff against another mutant (because always defecting is the unique best reply to itself). This implies that $\hat{q} = 1$ and, thus, $\eta \equiv d$.

Remark 5. Takahashi (2010) studies submodular Prisoner Dilemma games in a similar setup to ours[13] and proves that (I) they do not admit any strict equilibria; and (II) any individually rational payoff can be supported as the a sequential equilibrium payoff. Theorem 3 shows that defection is the unique stable outcome also when using the mild notion of weak stability. This implies that any equilibrium that induces a different outcome than always defecting must be dynamically unstable. Specifically, in Takahashi's equilibria each player is indifferent between cooperating and defecting after each observation. However, this implies that a group of mutants who always defect would strictly outperform the incumbents (regardless of the mutant's frequency): the mutants achieve the same payoff (relative to the incumbents) when being matched against an incumbent, and a higher payoff when being matched against a mutant (because always defecting is the unique best reply to itself).

6.3 Stable Cooperation (Observing Action Profiles)

In this section, we show that cooperation can be stable when players observe action profiles (rather than actions). Specifically, Theorem 4 shows that observation of a single action profile is enough to sustain cooperation as a limit strong evolutionarily stable outcome.

[13] Takahashi's (2010) setup differs in two aspects: (1) each agent can observe the entire history of past actions of his partner, and (2) agents can condition their behavior on their own past play (though, they do not do so in the characterized equilibria). We discuss the adaptation of our results to Takahashi's framework in Section 7.2.

Theorem 4. *Assume that G is a Prisoner's Dilemma game, the observation structure is $(0, p_2, 1)$ and $g < p_2 < 1$. Then there exists a limit evolutionarily stable strategy σ^* with a unique outcome $\eta^* \equiv c$.*

Sketch of Proof (see formal proof in Appendix A.4). Strategy σ^* includes a single policy s^* that induces players to cooperate unless they observe that the opponent was the sole defector in the past; in this case they defect. We consider a converging sequence of perturbed games in which players who tremble by choosing, with small probability, the policy that always defects. Note that all action profiles are observed with positive probability in these perturbed games. Moreover, when an incumbent is observed to play (d, c), it implies that he is a "trembler" who follows policy d and thus is going to defect in the current interaction. Thus, mutants who cooperate instead of defecting after observing (d, c) are strictly outperformed when facing the incumbents: they suffer an immediate loss of g, without gaining any indirect advantage from being observed to cooperate. Next, we show that mutants who always defect are strictly outperformed. Note that the probability that an incumbent defects when facing a mutant must be p_2 times the probability that an incumbent cooperates when facing a mutant. This is because an incumbent defects against a mutant iff he observes (with probability p_2) an action profile in which the partner (the mutant) defected and her past opponent (who is most likely an incumbent) cooperated. This implies that the incumbents defect with probability $\frac{p_2}{p_2+1}$ against the mutants. Thus the mutants' payoff against the incumbents is given by: $\frac{1+g}{p_2+1}$, which is strictly smaller than the incumbents' payoff among themselves (which is equal to one) if $p_2 > g$. This implies that if the mutants are sufficiently rare, they are strictly outperformed. □

Remark 6. The stability of cooperation in the proof above relies on a particular kind of trembles, in which tremblers are always defecting. The construction can be adapted to many other perturbations, as long as players have policy-trembles (possibly, in addition to action-trembles), such that observing a partner being the sole defector implies that he is more likely to defect again. This inference is strongest in the perturbation analyzed in the proof above. The inference may be weaker with other perturbations, and would thus require higher observation probability or/and observation of several action profiles, in order to sustain stable configuration. On the other hand, if the perturbations only include action-trembles, then stable cooperation cannot be sustained. In Section 7.1 we show that in a variant of the model in which the observed actions are public signals, then cooperation can be sustained also with action-trembles. The remark also implies (with minor adaptations) to theorem 5 below.

6.4 Stable Cooperation (Supermodular Prisoner's Dilemma)

In this section, we show that cooperation can be stable when players observe actions and the underlying game is supermodular. Specifically, Theorem 5 shows that observation of a single action (with a sufficient probability) is enough to sustain cooperation as a limit strong evolutionarily stable outcome.

Theorem 5. *Assume that G is a strictly supermodular Prisoner's Dilemma game (i.e. $l > g$), the observation structure is $(p_1, 0, 1)$ and $\frac{g}{1+g} < p_1$. Then there exists a limit evolutionarily stable strategy σ^* with a unique outcome $\eta^* \equiv c$.*

Sketch of Proof (see formal proof in Appendix A.5). Strategy σ^* includes two policies in its support: (1) s_t that plays "tit-for-tat" - defects iff observing a past defection of the partner; and (2) s_C that always cooperates.[14] The frequency of policy s_t, $0 < q < 1$, is defined such that both polices yield the same payoff in a population in which ϵ of the agents tremble and choose by mistake the policy of always defecting, while the non-trembling agents follow σ^* (see the explicit formula of this frequency as a function of l, g, p_1 in the Appendix A.5). We consider a converging sequence of perturbed games in which trembling players choose the policy that always defects.

When a player observes a past cooperation or a non-informative signal, the partner is most likely to be a non-trembling cooperative incumbent, and cooperation against him is a strict best reply (as the additional immediate g points obtained by defecting are outweighed by the indirect loss induced by making future opponents more likely to defect). When observing a past defection, the partner is more likely to be a trembling defector, and both actions are best replies against him (as the additional immediate gain is a mixed average of g and $l > g$, and q has been defined such that the future indirect loss is the exactly the same as the immediate gain). If the population is invaded by mutants who defect (on average) with higher (lower) probability than q when observing defection, then they are strictly outperformed: the aggregate probability that defecting today would induce a future opponent to defect is higher (lower) in the post-entry population relative to the pre-entry population, and as a result defecting after observing defection yields a lower payoff than cooperating. Finally, a polymorphic group of mutants who defect with probability q when observing defection yields a lower payoff than the incumbents because the supermodularity of the payoffs imply that the payoff of a policy as a function of its own defection probability after observing defection is strictly convex. □

Remark. Takahashi (2010) shows that if players observe the entire past history of play of their opponent, then they cooperation can be supported as the outcome of a strict equilibrium. Theo-

[14] As discussed in the formal proof in the appendix for low values of p_1 ($\frac{g}{1+g} < p_1 \leq \frac{l}{1+2\cdot l-g}$) strategy σ^* includes only the "tit-for-tat" policy.

rem 5 substantially strengthens this result by showing that that cooperation can stable also when each player only observes a single past action of the partner (though, it requires us to slightly weaken the solution concept from strict equilibrium to evolutionary stability).

7 Variants and Extensions

This section studies the robustness of our result in three variants and extensions of the model: (1) public signals, (2) endogenizing the observation structure, and (3) evolution of preferences.

7.1 Public Signals

In the main model we assume that the signal about the opponent's behavior is private. In some applications it might be more reasonable to assume that the signals are public. In particular, if we consider an online interaction between traders through an intermediary web site that publicly presents feedback about the past behavior of the traders (e.g., eBay), then in such interactions the signals about the past behavior (e.g. the trader's feedback summary) are public. Another setup in which public signals fit well is an environment in which a player observes the last k actions of the partner, rather than k random past actions. In such environments, the signals are essentially public because each player remembers his own recent history. In what follows we show how to adapt the model to public signals, and we analyze the influence on our results.

Changes to the model Before playing the game, each player *publicly* observes the signals about the opponent's past behavior. Because both players observe both signals, we redefine M as follows: $M = M_1 \times M_2 = \left(\phi \cup A^k \cup (A \times A)^k\right) \times \left(\phi \cup A^k \cup (A \times A)^k\right)$, where the first component is interpreted as the partner's observed past behavior and the second component as the player's own observed past behavior. We denote a typical element of M_1 as m_1 (the message about the partner's behavior), and a typical element of M_2 as m_2 (the message about the player's own behavior). All other details of the model remain the same.

Adaptations to the results It is relatively straightforward to show that our first two main results (Theorems 2-3) remain the same with public signals, and the proofs require only minor adaptations. In what follows we show that we can strengthen Theorem 4 in the setup of public signals, and obtain robust stable cooperation with respect to any sequence of action perturbations with full support (unlike the case of private signals, in which the stability relies on policy perturbations). Formally:

Theorem 6. *Assume that the underlying game is Prisoner's Dilemma, the observation structure is* $(0, p_2, 1)$ *with* $p_2 > g$*, and the signals are public. Then there exists an action-strict evolutionary stable strategy* σ^* *with a unique outcome* $\eta^* \equiv c$*.*

Sketch of proof (formal proof is in Appendix A.6). Strategy σ^* includes a single policy:

$$s^*(m) = \begin{cases} d & m_1 = (d,c) \text{ or } m_2 = (d,c) \\ c & \text{otherwise} \end{cases}.$$

That is, the policy induces players to cooperate unless they observe that either of the players was the sole defector in the past; in this case they defect. We consider an arbitrary converging sequence of action perturbations with full support (i.e., players rarely tremble and play the "wrong" action with small positive probability). Note that all action profiles are observed with positive probability in these perturbed games. Moreover, mutants who cooperate instead of defecting after observing (d,c) are strictly outperformed when facing the incumbents: they suffer an immediate loss of l, without gaining any indirect advantage from being observed to cooperate (as the opponent is going to defect with probability very close to one). The argument as to why mutants who always defect are strictly outperformed is the same as the one given after Theorem 4. This implies that if the mutants are sufficiently rare, they are strictly outperformed. □

Remark 7. The policy s^* is closely related to the strategy "Pavlov" (a.k.a., "win-stay, lose-change") in the standard repeated Prisoner's Dilemma in which a player defects iff the players played different actions in the previous round (see, Kraines & Kraines, 1989; Nowak & Sigmund, 1993, and Heller, 2015). "Pavlov" was used several times in the literature to support cooperation with direct reciprocity (players interact in repeated interactions, and each player rewards his past opponent's behavior), but to the best of our knowledge the current paper is the first to apply it to in a setup of indirect reciprocity (in which each pair of players meet only once).

7.2 Non-Stationary Policies

The main model implicit assumes that agents can only use stationary policies that are independent on both time and on the agent's history (his own past actions and the actions players by his past partners). In this section we sketch how to relax this assumption.

Changes to the model: We consider a continuum of agents who repeatedly play the prisoner's dilemma with varying partners in the same population. At each period $t = 1, 2, ...$ players are randomly matched into pairs and play the underlying game. At the first stage each player

obtains a non-informative signal about his partner. At each of the subsequent stages, each player privately observes either: (1) k independent draws (with replacement) of his partner's past actions with probability p_1, (2) k independent draws (with replacement) of his partner's (and his past opponent) past action-profiles with probability p_2, and (3) a non informative signal with the remaining probability $1 - p_1 - p_2$. Each player's total payoff is a discounted sum of the stage payoffs with a common discount factor $\delta \in (0, 1)$.

A private history (information set) is a tuple $\left(\tau, ((a_t, a'_t))_{1 \leq t < \tau}, (m_t)_{2 \leq t \leq \tau}\right)$, where $\tau \in \mathbb{N}$ denotes the period, a_t is the agent's own action at past period $t < \tau$, $a_{t'}$ is the past partner's action at period $t < \tau$, and $m_t \in M$ is the signal observed about the partner at round t (m_τ is the signal observed about the current partner's past). A (non-stationary) policy of player is a mapping that assigns a mixed action to each private history. A strategy (or population state) is a finite-support distribution over the set of policies. The adaptation of the remaining aspects of the model and of the solution concepts is straightforward.

Adaptation of the results: The main drawback of this approach is the large set of policies, which makes the uniqueness analysis intractable. We conjecture that Theorem 3 (i.e. only defection is limit evolutionary stable when observing actions in a submodular prisoner dilemma game) remains true in this setup, but we leave the proof for future research.

All the other results can be adapted to this setup in a relatively straightforward way:

- Theorem 2 (any strict equilibrium is evolutionary stable): The policy that plays a symmetric strict equilibrium of the underlying stage game regardless of the history, is the unique strict best-reply to itself in any slightly-perturbed environment with full-support action trembles, and as a result it is an action-strict limit evolutionary stable for any discount factor.

- Theorems 4-5 (stable cooperation when either observing an action-profile or in a supermodular prisoner dilemma): By minor adaptations to the proofs, one can show that there exists $\bar{\delta} < 1$, such that for each $\bar{\delta} < \delta < 1$, the policies in the proofs remain stable also when non-stationary policies are allowed.

7.3 Endogenous Observability

In this subsection we extend our analysis to a setup in which the observation probabilities are endogenously determined by the players.

Changes to the model Each individual in the population is characterized by a *type* which is a pair (e, s), where $e \in \mathbb{N}$ is the effort level, and s is the policy. The fixed observation probabilities p_1

and p_2 in the basic model are replaced with weakly increasing observability functions $p_1, p_2 : \mathbb{N} \to [0,1)$, satisfying $p_1(n) + p_2(n) \leq 1$ for each n. These functions have the following interpretation: an individual who spends effort e observes k past actions (action profiles) of his opponent with probability $p_1(e)$ ($p_2(e)$). Let $T = \mathbb{N} \times S$ be the set of all types (pairs of effort and policy), with typical element t.[15] Given type t, let $e(t)$ ($s(t)$) denote its effort (policy). We redefine a strategy to be a distribution (with a finite support) over T. The definitions of outcomes, consistent outcomes, and configurations remain essentially the same except that the probabilities p_1 and p_2 are replaced with their function counterparts. Finally, the expression for the payoff of a type is modified by subtracting the cost of the effort from the payoff, i.e.,

$$\pi_t(\sigma, \eta) = \sum_{(a,a') \in A \times A} \pi(a, a') \cdot \mu_{t,\sigma,\eta}(a, a') - c(e(t)),$$

where $c : \mathbb{N} \to \mathbb{R}^+$ is a strictly increasing cost function. The definitions of evolutionary, weak and limit stability remain the same (except the redefinition of strategies and the reduction of the effort cost as described above).

Adaptations of the results Recall that Theorems 2-3 hold for all observation structures and Theorem 4 holds if $p_2 > g$.. It is relatively straightforward to show that these results also in this extended setup with minor adaptations to the proofs (Theorem 4 hold if $p_2(0) > g$).

7.4 Evolution of Preferences

In this subsection we extend the model to analyze the evolution of subjective preferences.

Changes to the model Let $\Theta = [0,1]^{|A|^2}$ be the set of all feasible subjective *preferences* - utility functions on $A \times A$. A configuration with preferences is a triple consisting of a distribution of preferences, a policy for each preference, and a consistent outcome, satisfying that each policy is a subjective best-reply (i.e, a Bayesian Nash equilibrium given the subjective preferences).

Definition 13. A *configuration with preferences* is a triple (μ, S_μ, η) where:

1. $\mu \in \Delta(\Theta)$ is a distribution (with a finite support) over the set of types.
2. $S_\mu = (s_\theta)_{\theta \in C(\mu)}$ is a profile of policies, where each s_θ is the policy of preference $\theta \in \Theta$. Let $O_{C(\mu)}$ be the set of outcomes $\eta : C(\mu) \times C(\mu) \to \Delta(A)$ that describe the behavior of each preference conditional on being matched with each other preference. The pair (μ, S_μ) defines a dynamic mapping $f_{(\mu,S_\mu)}$ between outcomes in $O_{C(\mu)}$ (like f_σ in Section 2).

[15] The results are similar with with a continuum of feasible efforts.

3. $\eta \in O_{C(\mu)}$ is a consistent outcome with respect to (μ, S_μ) (i.e., $f_{(\mu,S_\mu)}(\eta) = \eta$).
Let m be a signal that is observed with positive probability given outcome η. Let $\eta_{(\mu,S_\mu,\eta),\theta,m} \in \Delta(A)$ be the average mixed action of a random partner conditional on being matched with a player with preference θ, and on that latter player observing signal m.

4. We require each policy to be a subjective best reply; that is, for each $\theta \in C(\mu)$, each message m (with positive probability), and each action a:

$$\theta\left(s_\theta(m), \eta_{(\mu,S_\mu,\eta),\theta,m}\right) \geq \theta\left(a, \eta_{(\mu,S_\mu,\eta),\theta,m}\right).$$

Given two distributions $\mu^*, \mu' \in \Delta(\Theta)$ and $\epsilon > 0$, let $\mu_\epsilon = \mu_{\mu^*,\epsilon,\mu'}$ be the mixture distribution: $\mu_\epsilon = (1-\epsilon)\cdot\mu^* + \epsilon\cdot\mu'$, and let a post-entry configuration be a configuration with preferences where the first component is a mixture distribution. Finally, we say that the post-entry configuration $(\mu_\epsilon, S_{\mu_\epsilon}, \eta_\epsilon)$ is *focal* with respect to (μ, S_μ, η) if the incumbents' behavior among themselves remain unchanged in the post-entry configuration: for each $\theta, \theta' \in C(\mu)$, $S_{\mu_\epsilon}(\theta) = S_\mu(\theta)$ and $\eta_\theta(\theta') = (\eta_\epsilon)_\theta(\theta')$ (see the related notion of focality and its motivation in Dekel *et al.*, 2007). Next. we define a strong notion of evolutionary stability, which requires the behavior in any post-entry configuration to be focal , and any group of mutant preferences to be outperformed (weakly outperformed if they behave the same as the incumbents, and strictly outperformed otherwise).

Definition 14. A configuration with preferences $(\mu^*, S_{\mu^*}, \eta^*)$ is *evolutionary stable* if it is balanced, and there exists $\bar{\epsilon} > 0$ such that for each $\epsilon \in (0, \bar{\epsilon})$, and each mutant distribution of preferences $\mu' \in \Delta(\Theta)$ and each post-entry configuration $(\mu_\epsilon, S_{\mu_\epsilon}, \eta_\epsilon)$: (1) the post-entry configuration is focal, (2) the mutants are outperformed: $\pi_{\mu'}(\mu_\epsilon, S_{\mu_\epsilon}, \eta_\epsilon) \leq \pi_{\mu^*}(\mu_\epsilon, S_{\mu_\epsilon}, \eta_\epsilon)$, with a strict inequality if the post-entry behavior is different, i.e., if μ_ϵ induces a different aggregate distribution over action profiles than μ^*.

Adaptation to Theorem 2 Finally, we extend Theorem 2 to this setup, and show that strict equilibrium of the underlying game is evolutionary stable for any observation structure also in this setup. This contradicts the main stylized result in the literature of the evolution of preferences that only efficient outcomes may be stable if the observability probability is sufficiently high. This is formalized as follows. Let θ_a denote a preference in which action a is strictly dominant.

Proposition 1. *If a^* is a strict equilibrium of the underlying game, then the configuration with preferences (θ_{a^*}, a^*, a^*) is evolutionary stable.*

The proof is analogous to Theorem 2's proof, and is thus omitted. We leave for future research

the presentation of a weaker notion of stability for uniqueness results, and the extension of our other result to the setup with preferences.

Relations with the existing literature Several papers have studied the evolution of preferences using the so-called "indirect evolutionary approach" (see, e.g., Güth & Yaari 1992; Ok & Vega-Redondo 2001; Sethi & Somanathan 2001; Dekel *et al.* 2007; Heifetz *et al.* 2007; Herold & Kuzmics 2009). These papers consider agents that are endowed with subjective preferences that may differ from the material payoffs. With positive probability each agent observes the subjective preferences of the opponent, and the agents play a Bayesian-Nash equilibrium induced by their subjective preferences and their signals. In a stable state, all incumbent preferences achieve the same fitness payoff, and any sufficiently small group of mutants with different subjective preferences is outperformed.

The standard argument for observing preferences is that people give signals that provide clues as to their feeling (e.g., a blush that may reveal a lie). As discussed in Robson & Samuelson (2010, Section 2.5), the emission of such signals and their correlation with preferences are themselves the product of evolution, and thus it is not clear what prevents the appearance of a mimic who emits the signal without having the associated preferences. Robson & Samuelson (2010) summarize their discussion by suggesting that "the indirect evolutionary approach will remain incomplete until the evolution of preferences, the evolution of signals about preferences, and the evolution of reactions to these signals, are all analyzed within the model."

Our model share many aspects of the existing literature, with one key difference: players observe past behavior and infer about the subjective preferences (a "revealed preferences" approach), rather then directly observe the partner's preferences. This key difference allows our model to answer Robson & Samuelson (2010)'s criticism. Another issue with the existing literature is that it assumes that two preferences that slightly differ from each other (e.g., slightly different numbers in the payoff matrix, without any real influence on the best-reply correspondence) induce completely different signals. Due to this issue, many of the results in that literature crucially depend on the influence of non-generic subjective preferences, such as agents that are completely indifferent between all the action profiles (e.g., Dekel *et al.* 2007, Prop. 2). This dependency on non-generic preferences is absent from our setup.

8 Conclusion

We study a setup in which players are randomly matched, and each player may observe signals about the opponent's past behavior. We present three main results: (1) strict equilibria of the

underlying game are stable outcomes, (2) defection is the unique stable outcome in the Prisoner's Dilemma when players only observe past actions, and (3) cooperation can be a stable outcome if players observe action profiles (and the stability becomes more robust if the signals are public). Moreover, we show that our model can be extended to study endogenous observability (which depends on the players' efforts) and the evolution of subjective preferences.

Throughout the paper we interpreted the players as naive agents who just follow their programmed strategies. However, similar to other applications of evolutionary stability a la Maynard Smith & Price (1973), the results are robust to the presence of sophisticated agents who explicitly maximize their payoff. This is illustrated in the non-stationary variant (Section 7.2), in which any limit weakly stable strategy is is a symmetric Nash equilibrium of the dynamic game.

Directions for Future Research In what follows we sketch four interesting directions for future research. We mainly applied our model to the Prisoner's Dilemma. It will be interesting to apply the model to other games, and in particular, to the Hawk-Dove game, in which we conjecture that the environment admits a stable heterogeneous population in which "committed Hawks" and "flexible players" co-exist.

Second, our extension to subjective preferences is somewhat limited because players only interact in single game. It seems intriguing, to study richer environments in which players are endowed with "universal" (non-game-specific) preferences over fitness profiles, and they interact in different games (and use the same preferences in all these games). Another interesting direction (pursued in a companion working paper, Heller & Mohlin, 2014) is allowing agents to spend effort in deception - influencing the signal observed by the opponent.

Third, our model assumes that players directly observe past actions of the partner. In many economic setups, it seems more plausible that agents only observe the reports of other agents about the past interactions of their partner (e.g., the trader's feedback profile in eBay). A policy in this setup should specify both the played actions, as well as the reports to other agents, and it will be interesting to characterize stable outcomes in such setups.

Finally, some important interactions may be better modeled as asymmetric games between separate populations (e.g., a population of consumers and a population of professional sellers interacting in on-line trade like Amazon). It will be interesting to extend our methodology to this setup, and to study stable outcomes in such multiple-population interactions.

Additional Related Literature A few papers study how cooperation can be supported in the Prisoner's Dilemma without observing the opponent's past behavior. Ellison (1994) (extending previous results of Kandori, 1992 and Harrington Jr, 1995) shows how cooperation can be

supported in finite populations by "contagious" punishments (if one player defects at stage t, his partner defects from period $t+1$, infecting another player who defects from period $t+2$ on, etc); however, such "contagious" punishments can only work in finite populations, and if the players are sufficiently patient with respect to the population size. Fujiwara-Greve & Okuno-Fujiwara (2009) show how cooperation can be supported in "voluntarily separable" repeated Prisoner's Dilemma, in which each player can unilaterally end and start with a randomly assigned new partner with no information flow. van Veelen *et al.* (2012) show how cooperation can be stable in structured populations in which agents are more likely to interact with similar partners.

Rosenthal (1979) presented an early model in which players in a population are randomly matched, and each player can observe his opponent's last action (See also Okuno-Fujiwara & Postlewaite, 1995). A few papers constructed models in which players may observe the partner's planned action (rather than past action): Robson (1994) presented a model in which each player has small probability to observe his partner's planned action and revise his own action; Solan & Yariv (2004) dealt with a setup in which one of the players can spend effort and observe his opponent's action before playing.

A Proofs

A.1 Proof of Theorem 1 ("2" Implies "1")

In this section we present the formal proof "2" implies "1" in Prop. 1. That is, we assume that $k \cdot p_1 + 2 \cdot k \cdot p_2 \leq 1$ and $p_1 < 1$ and show that every strategy in E admits a unique consistent outcome. Let σ be a strategy, and let $\eta, \eta' \in O_{c(\sigma)}$ be two consistent outcomes. In order to shorten the notation in this section, we omit the subscript σ in the remainder of the proof. In what follows we show that f (i.e. f_σ) is a contraction mapping (which implies that $\eta = \eta'$).

A.1.1 Notation and Definitions

We need to introduce some additional notation for this proof. Define $\eta^k : C(\sigma) \to \triangle\left(A^k\right)$ as a mapping that assigns for each policy s the probability distribution over k-tuples of actions that a player observes conditional on being matched with an opponent with policy s. Formally:

$$\eta^k(s)\left((a_i)_{i\leq k}\right) := \eta_s^k\left((a_i)_{i\leq k}\right) = \prod_{i\leq k}\left(\eta_s(a_i)\right) = \prod_{i\leq k}\left(\sum_{s'\in C(\sigma)}\sigma(s')\cdot\eta_{s'}(s')(a_i)\right).$$

Similarly, define $\psi^k_{\sigma,\eta} : C(\sigma) \to \triangle\left((A \times A)^k\right)$ as a mapping that assigns for each policy s the probability distribution over k-tuples of action-profiles (i.e. probability to each k-tuple $(a_i, a'_i)_{i \leq k}$) that a player observes conditional on being matched with an opponent with policy s. Formally:

$$\psi^k_{s,\eta}\left((a_i, a'_i)_{i \leq k}\right) = \prod_{i \leq k} \left(\psi_{s,\eta}(a_i, a'_i)\right) = \prod_{i \leq k} \left(\sum_{s' \in C(\sigma)} \sigma(s') \cdot \eta_s(s')(a_i) \cdot \eta_{s'}(s)(a'_i) \right).$$

Define η'^k and $\psi'^k_\eta(s)$ in an analogous way. Note that $\eta^1_s \equiv \eta_{s,\sigma}$ and $\psi^1_s \equiv \psi_{s,\sigma,\eta}$, where the notions in the right hand side are as defined in Section 3.1.

We measure distance between *finite support probability distributions* with the L_1-norm as follows: let X be a finite set and $\Delta(X)$ the set of probability distributions on X. Given two probability distributions $\xi, \xi' \in \Delta(X)$, their distance is defined as the sum of the absolute differences in the weights they assign to the different elements of X:

$$\|\xi - \xi'\|_1 = \sum_{x \in X} |\xi(x) - \xi'(x)|.$$

For example,

$$\left\|\eta^k_s - \eta'^k{}_s\right\|_1 = \sum_{(a_i)_{i \leq k} \in A^k} \left|\eta^k_s\left((a_i)_{i \leq k}\right) - \eta'^k{}_s\left((a_i)_{i \leq k}\right)\right|, \text{ and}$$

$$\left\|\psi^k_{s,\eta} - \psi^k_{s,\eta'}\right\|_1 = \sum_{(a_i,a'_i)_{i \leq k} \in A \times A} \left|\psi^k_{s,\eta}\left((a_i, a'_i)_{i \leq k}\right) - \psi^k_{s,\eta'}\left((a_i, a'_i)_{i \leq k}\right)\right|.$$

We measure distance between *sets of finite support probability distributions with the same support* with the help of the L_∞-norm. For any two sets of distributions $\gamma, \gamma' \subseteq \Delta(X)$, we define

$$\|\gamma - \gamma'\|_\infty = \max_{\xi \in \gamma, \xi' \in \gamma'} \|\xi - \xi'\|_1.$$

For example, $\left\|\eta^k - \eta'^k\right\|_\infty = \max_{s \in C(\sigma)} \left\|\eta^k_s - \eta'^k_s\right\|_1$, and $\left\|\psi^k_\eta - \psi^k_{\eta'}\right\|_\infty = \max_{s' \in C(\sigma)} \left\|\psi^k_{s,\eta} - \psi^k_{s,\eta'}\right\|_1$.

Finally, we also use the L_∞-norm to measure the distance between η_s and η'_s (since η_s can be interpreted as representing the set $\{\eta_s(s')\}_{s' \in c(\sigma)}$):

$$\|\eta_s - \eta'_s\|_\infty = \max_{s' \in C(\sigma)} \|\eta_s(s') - \eta'_s(s')\|_1,$$

and we use two notions of norms $\|\|_{\infty,\infty}$ and $\|\|_{\infty,1}$ to measure distances between η and η':

$$\|\eta - \eta'\|_{\infty,\infty} = \max_{s \in C(\sigma)} \|\eta_s - \eta'_s\|_\infty, \quad \|\eta - \eta'\|_{\infty,1} = \max_{s \in C(\sigma)} \|\eta_s - \eta'_s\|_1.$$

Observe that $\|\eta_s - \eta'_s\|_1 \leq \|\eta_s - \eta'_s\|_\infty$ and $\|\eta - \eta'\|_{\infty,1} \leq \|\eta - \eta'\|_{\infty,\infty}$.

A.1.2 Bounding the Distance Between Observed Action Profiles

We begin by showing that the distance between the distributions of an observed action profile is at most twice the distance between the distributions of an observed action.

$$
\begin{aligned}
&\|\psi_{s,\eta} - \psi_{s,\eta'}\|_1 = \sum_{(a,a')\in A^2} |\psi_{s,\eta}(a,a') - \psi_{s,\eta'}(a,a')| \\
= &\sum_{(a,a')\in A^2} \left| \sum_{s'\in C(\sigma)} \sigma(s') \cdot (\eta_s(s')(a) \cdot \eta_{s'}(s)(a') - \eta'_s(s')(a) \cdot \eta'_{s'}(s)(a')) \right| \\
\leq &\sum_{(a,a')\in A^2} \sum_{s'\in C(\sigma)} \sigma(s') |(\eta_s(s')(a) \cdot \eta_{s'}(s)(a') - \eta'_s(s')(a) \cdot \eta'_{s'}(s)(a'))| \\
= &\sum_{s'\in C(\sigma)} \sigma(s') \cdot \sum_{(a,a')\in A^2} |(\eta_s(s')(a) \cdot \eta_{s'}(s)(a') - \eta'_s(s')(a) \cdot \eta'_{s'}(s)(a'))| \\
= &\sum_{s'\in C(\sigma)} \sigma(s') \cdot \sum_{(a,a')} |(\eta_s(s')(a) \cdot (\eta_{s'}(s)(a') - \eta'_{s'}(s)(a')) + \eta'_{s'}(s)(a')(\eta_s(s')(a) - \eta'_{s'}(s')(a)))| \\
< &\sum_{s'\in C(\sigma)} \sigma(s') \cdot \sum_{(a,a')} (\eta_s(s')(a) \cdot |\eta_{s'}(s)(a') - \eta'_{s'}(s)(a')| + \eta'_{s'}(s)(a') \cdot |\eta_s(s')(a) - \eta'_{s'}(s')(a)|) \\
= &\sum_{s'\in C(\sigma)} \sigma(s') \cdot \sum_{a'\in A} |\eta_{s'}(s)(a') - \eta'_{s'}(s)(a')| + \sum_{a\in A} |\eta_s(s')(a) - \eta'_{s'}(s')(a)| \\
= &\sum_{s'\in C(\sigma)} \sigma(s') \cdot (\|\eta_{s'}(s) - \eta'_{s'}(s)\|_1 + \|\eta_s(s') - \eta'_{s'}(s')\|_1) \leq \sum_{s'\in C(\sigma)} \sigma(s') \\
&\cdot (\|\eta_{s'} - \eta'_{s'}\|_\infty + \|\eta_s - \eta'_s\|_\infty) \leq \sum_{s'\in C(\sigma)} \sigma(s') \cdot \left(\|\eta - \eta'\|_{\infty,\infty} + \|\eta - \eta'\|_{\infty,\infty} \right) = 2 \cdot \|\eta - \eta'\|_{\infty,\infty}.
\end{aligned}
$$

The second inequality is strict because the elements inside the "||" in the l.h.s. of the strict inequality include both positive and negative elements.

A.1.3 Bounding the Distance Between Observed Tuples

Let $k \geq 2$. Next we show that the distance between the distribution of tuples of k observations is at most k times the distance between a single observed action. This is done as follows:

$$
\begin{aligned}
\left\| \eta_s^k - {\eta'}^k{}_s \right\|_1 &= \sum_{(a_i)_{i\leq k}\in A^k} \left| \eta_s^k\left((a_i)_{i\leq k}\right) - {\eta'}^k{}_s\left((a_i)_{i\leq k}\right) \right| \\
&= \sum_{(a_i)_{i\leq k}\in A^k} \left| \prod_{i\leq k} \eta_s(a_i) - \prod_{i\leq k} \eta'_s(a_i) \right|
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{(a_i)_{i\leq k}\in A^k}\left|\sum_{i\leq k}\left(\eta_s\left(a_i\right)-\eta'_s\left(a_i\right)\right)\cdot\prod_{j>i}\eta_s\left(a_j\right)\cdot\prod_{j<i}\eta'_s\left(a_j\right)\right| \\
&< \sum_{(a_i)_{i\leq k}\in A^k}\sum_{i\leq k}\left(\left|\eta_s\left(a_i\right)-\eta'_s\left(a_i\right)\right|\cdot\prod_{j>i}\eta_s\left(a_j\right)\cdot\prod_{j<i}\eta'_s\left(a_j\right)\right) \\
&= \sum_{i\leq k}\left(\sum_{a\in A}\left|\eta_s\left(a\right)-\eta'_s\left(a\right)\right|\cdot\left(\sum_{(a_j)_{j>i}\in A^{n-i}}\prod_{j>i}\eta_s\left(a_j\right)\right)\cdot\left(\sum_{(a_j)_{j<i}\in A^{i-1}}\prod_{j<i}\eta'_s\left(a_j\right)\right)\right) \\
&= \sum_{i\leq k}\left(\sum_{a\in A}\left|\eta_s\left(a\right)-\eta'_s\left(a\right)\right|\cdot 1\cdot 1\right)=k\cdot\sum_{a\in A}\left|\eta_s\left(a\right)-\eta'_s\left(a\right)\right| \\
&= k\cdot\ \left\|\eta_s-\eta'_s\right\|_1\leq k\cdot\left\|\eta-\eta'\right\|_{\infty,1}\leq k\cdot\left\|\eta-\eta'\right\|_{\infty,\infty}.
\end{aligned}
$$

The first equality is due to the independence of different observations. The third equality is implied by the fact the sum in the third row includes many elements that cancel out (appearing once with a positive sign and once with a negative sign), and the only elements that do not cancel out are those appearing in the second row. The inequality is strict because the set of numbers inside the "||" in the l.h.s. of the inequality include both positive and negative elements. The first equality on the penultimate row holds because each sum adds the probabilities of disjoint and exhausting events.

An analogous argument yields the same result for observed tuples of action profiles:

$$
\left\|\psi^k_{s,\eta}-\psi^k_{s,\eta'}\right\|_1<k\cdot\left\|\psi_{s,\eta}-\psi_{s,\eta'}\right\|_1\leq k\cdot 2\cdot\left\|\eta-\eta'\right\|_{\infty,\infty},
$$

with a strict inequality if $k>1$.

A.1.4 Showing that $f(\eta)$ is a Contraction Mapping

We bound the distance between $\left(f\left(\eta\right)\right)_s\left(s'\right)$ and $\left(f\left(\eta'\right)\right)_s\left(s'\right)$ as follows.

$$
\begin{aligned}
&\left\|\left(f\left(\eta\right)\right)_s\left(s'\right)-\left(f\left(\eta'\right)\right)_s\left(s'\right)\right\|_1=\sum_{a\in A}\left|\left(f\left(\eta\right)\right)_s\left(s'\right)\left(a\right)-\left(f\left(\eta'\right)\right)_s\left(s'\right)\left(a\right)\right| \\
= \sum_{a\in A}\left|\begin{array}{c} p_1\cdot\sum_{(a_i)_{i\leq k}\in A^k}\left[\prod_{i\leq k}\eta_{s'}\left(a_i\right)\cdot s\left(\left(a_i\right)_{i\leq k}\right)\left(a\right)-\prod_{i\leq k}\eta'_{s'}\left(a_i\right)\cdot s\left(\left(a_i\right)_{i\leq k}\right)\left(a\right)\right]+p_2\cdot \\ \sum_{\left(a_i,a'_i\right)_{i\leq k}}\left[\left(\prod\psi_{s',\eta}\left(a_i,a'_i\right)\cdot s\left(\left(a_i,a'_i\right)_{i\leq k}\right)\left(a\right)-\prod_{i\leq k}\psi_{s',\eta'}\left(a_i,a'_i\right)\cdot s\left(\left(a_i,a'_i\right)_{i\leq k}\right)\left(a\right)\right)\right]\end{array}\right| \\
= \sum_{a\in A}\left|\begin{array}{c} p_1\sum_{(a_i)_{i\leq k}\in A^k}s\left(\left(a_i\right)_{i\leq k}\right)\left(a\right)\cdot\left[\prod_{i\leq k}\eta_{s'}\left(a_i\right)-\prod_{i\leq k}\eta'_{s'}\left(a_i\right)\right] \\ +p_2\cdot\sum_{\left(a_i,a'_i\right)_{i\leq k}}s\left(\left(a_i,a'_i\right)_{i\leq k}\right)\left(a\right)\cdot\left[\left(\prod\psi_{s',\eta}\left(a_i,a'_i\right)-\prod_{i\leq k}\psi_{s',\eta'}\left(a_i,a'_i\right)\right)\right]\end{array}\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{a\in A}\left|p_1 \sum_{(a_i)_{i\leq k}\in A^k} s\left((a_i)_{i\leq k}\right)(a)\cdot\left[\prod_{i\leq k}\eta_{s'}(a_i)-\prod_{i\leq k}\eta'_{s'}(a_i)\right]\right| \\
&\quad+\sum_{a\in A}\left|p_2 \sum_{\left(a_i,a'_i\right)_{i\leq k}\in(A\times A)^k} s\left((a_i,a'_i)_{i\leq k}\right)\cdot\left[\prod\psi_{s',\eta}(a_i,a'_i)(a)-\prod_{i\leq k}\psi_{s',\eta'}(a_i,a'_i)\right]\right| \\
&\leq \sum_{a\in A}\sum_{(a_i)_{i\leq k}\in A^k} s\left((a_i)_{i\leq k}\right)(a)\cdot\left|p_1\cdot\prod_{i\leq k}\eta_{s'}(a_i)-p_1\cdot\prod_{i\leq k}\eta'_{s'}(a_i)\right| \\
&\quad+\sum_{a\in A}\sum_{\left(a_i,a'_i\right)_{i\leq k}\in(A\times A)^k} s\left((a_i,a'_i)_{i\leq k}\right)(a)\cdot\left|p_2\prod\psi_{s',\eta}(a_i,a'_i)-p_2\prod_{i\leq k}\psi_{s',\eta'}(a_i,a'_i)\right| \\
&= \sum_{(a_i)_{i\leq k}} p_1\cdot\left|\prod_{i\leq k}\eta_{s'}(a_i)-p_1\prod_{i\leq k}\eta'_{s'}(a_i)\right|+\sum_{\left(a_i,a'_i\right)_{i\leq k}} p_2\cdot\left|\prod_{i\leq k}\psi_{s',\eta}(a_i,a'_i)-p_2\prod_{i\leq k}\psi_{s',\eta'}(a_i,a'_i)\right| \\
&= p_1\cdot\left\|\eta^k_{s'}-\eta'^k_{s'}\right\|_1+p_2\cdot\left\|\psi^k_{s',\eta}-\psi^k_{s',\eta'}\right\|_1\leq p_1\cdot k\cdot\|\eta-\eta'\|_{\infty,\infty}+p_2\cdot k\cdot 2\|\eta-\eta'\|_{\infty,\infty} \\
&= (p_1\cdot k+2\cdot p_2\cdot k)\cdot\|\eta-\eta'\|_{\infty,\infty}\leq\|\eta-\eta'\|_{\infty,\infty}.
\end{aligned}
$$

Moreover, the penultimate inequality is strict if either (i) $k>1$ or (ii) $p_2>0$. The last inequality is strict if (iii) $p_1\cdot k+2\cdot p_2\cdot k<1$. Observe, that at least one of the inequalities (i-iii) must hold if $k\cdot p_1+2\cdot k\cdot p_2\leq 1$ and $p_1<1$. Thus, we obtain the following strict inequality:

$$
\|f(\eta)-f(\eta')\|_{\infty,\infty}\leq\max_{s,s'\in C(\sigma)}\|(f_\sigma(\eta))_s(s')-(f_\sigma(\eta'))_s(s')\|_1 \quad<\quad \|\eta-\eta'\|_{\infty,\infty},
$$

which implies that f is a contraction mapping and $\eta=\eta'$.

A.2 Proof of Theorem 2 (Strict Equilibrium is Evolutionarily Stable)

Proof. Assume that a^* is a strict equilibrium of G. We have to show that strategy $\sigma^*\equiv a^*$ is an action-strict evolutionary stable. Observe first that σ^* admits a unique consistent outcome $\eta^*\equiv a^*$, and that the configuration (a^*,a^*) is balanced. This shows part (i) in Definition 11 of action-strict evolutionary stability.

Let $(\xi_n)_{n\in\mathbb{N}}$ be a converging sequence of full support action perturbations (i.e., $\lim_{n\to\infty}M(\xi_n)\to 0$). We construct a sequence of configurations $(\sigma_n,\eta_n)_{n\in\mathbb{N}}$ as follows: for each n, let σ_n be the strategy that puts all probability on the policy s_n, i.e. $\sigma_n\equiv s_n$, defined by

$$
s_n(\cdot)(a)=\begin{cases}\xi_n(a) & \text{if } a\neq a^*\\ 1-\sum_{a'\neq a^*}\xi_n(a') & \text{if } a=a^*\end{cases}.
$$

That is, the policy chooses a^* with the maximal allowed probability regardless of the signal about the opponent. Since actions are independent of signals there is a unique outcome η_n that is consistent with σ_n. Observe that $L(\zeta_n) \to 0$, each σ_n is feasible in $E(\xi_n)$, and the sequence of configurations (s_n, η_n) converge to (a^*, a^*). Thus we are left to show part 2(C) of Definition 11. That is, we have to show that $\exists \bar{\epsilon}, n_0 > 0$ such that $\bar{\epsilon}$ is an invasion barrier for σ_n for each $n \geq n_0$.

We will need a few pieces of notation. Let $l(a^*) > 0$ be the minimal loss from playing a different action against a^*, and let $g(a^*) \geq 0$ be the maximal gain that can be achieved by playing a different action profile:

$$l(a^*) = \min_{a \neq a^*} (\pi(a^*, a^*) - \pi(a, a^*)), \quad g(a^*) = \max_{(a,a') \in A \times A} (\pi(a, a') - \pi(a^*, a^*)).$$

We have:

$$\psi_n \equiv \psi_{s_n, \sigma_n, \eta_n}(a, a') = \begin{cases} \xi_n(a) \cdot \xi_n(a') & \text{if } a, a' \neq a^* \\ \left(1 - \sum_{a'' \neq a^*} \xi_n(a'')\right)^2 & \text{if } a = a' = a^*. \\ \left(1 - \sum_{a'' \neq a^*} \xi_n(a'')\right) \cdot \xi_n(a') & \text{if } a = a^* \neq a' \end{cases}$$

Pick a sufficiently small $\bar{\epsilon} < 1$ such that:

$$(1 - \bar{\epsilon}) \cdot 0.5 \cdot l(a^*) - \frac{\bar{\epsilon} \cdot g(a^*) \cdot (k+1)}{1 - 2 \cdot k \cdot \bar{\epsilon}} > 0. \tag{1}$$

Pick any$\epsilon \in (0, \bar{\epsilon})$. Let $\sigma' \in \Sigma(\zeta_n)$ be a mutant strategy and let $(\sigma_\epsilon, \tilde{\eta})$ be a post-entry configuration (with respect to the pre-entry strategy s_n . It remains to show that the incumbents outperform the mutants. Let q_0 be the probability that an incumbent plays an action different from a^* at each round:

$$q_0 = 1 - \sum_{a'' \neq a^*} \xi_n(a'').$$

Let q_l be the probability that a mutant plays $a \neq a^*$ when being matched with an incumbent:

$$q_l = 1 - \sum_{s' \in C(\sigma')} \sigma'(s') \cdot (\tilde{\eta}_n)_{s', \sigma_n}(a^*).$$

Note that for each signal m the strategy s_n is the unique strategy that minimizes the probability of playing $a \neq a^*$ after observing m. The fact that all signals are observed with positive probability (because the perturbation has full support) implies that $q_0 < q_l$. Let $q_g \in [0, 1]$ be the probability

that a mutant plays action $a \neq a^*$ when being matched against a mutant.

$$q_g = 1 - \sum_{s' \in C(\sigma')} \sigma'(s') \cdot (\tilde{\eta}_n)_{s',\sigma'}(a^*).$$

The consistency of outcome $\tilde{\eta}$ with the strategy σ_ϵ implies an upper bound on q_g as follows. Consider a mutant player (he) that is matched with a mutant partner (she). The partner can be distinguished from an incumbent only to the extent that the induced distribution of observed signals differ from the incumbents' distribution. That is, the L_1-distance between the mixed action played against an incumbent and the one played against a mutant is bounded by the L_1-distance between their induced distribution of signals. In each match, she plays a different action (than an incumbent) with probability of at most $q_g - q_0$ when being matched with a mutant partner (which happens with probability ϵ), and with probability of at most $q_l - q_0$ when being matched with an incumbent partner (which happens with probability $1-\epsilon$). When observing a match of the opponent with a mutant partner, the player may observe a different (non-incumbent) behavior also due to the action of the partner, which may play differently than the incumbents with probability of at most $q_g - q_0$. Thus, a bound on the L_1-distance between the distribution of signals induces by a mutant opponent and an incumbent opponent is : $k \cdot (\epsilon \cdot 2 \cdot (q_g - q_0) + (1-\epsilon) \cdot (q_l - q_0))$.

This yields the following upper bound on q_g:

$$\begin{gathered}
q_g - q_0 \leq k \cdot (\epsilon \cdot 2 \cdot (q_g - q_0) + (1-\epsilon) \cdot (q_l - q_0)) + (q_l - q_0) \Leftrightarrow \\
(q_g - q_0) \cdot (1 - k \cdot \epsilon \cdot 2) \leq (k \cdot (1-\epsilon) \cdot (q_l - q_0)) + (q_l - q_0) \Rightarrow \\
(q_g - q_0) \cdot (1 - k \cdot \epsilon \cdot 2) \leq (k+1) \cdot (q_l - q_0) \Leftrightarrow (q_g - q_0) \leq \tfrac{k+1}{1-2 \cdot k \cdot \epsilon} \cdot (q_l - q_0)
\end{gathered}$$

For a sufficiently large n_0 the trembles are sufficiently small such that, for each $n \geq n_0$ a mutant yields a loss of at least $\frac{l(a^*)}{2}$ when playing differently than the incumbents against an incumbent. Thus the difference between the incumbent's payoff and the mutant's payoff is at least:

$$\pi_{s_n}(\sigma_{s_n,\epsilon,\sigma'}, \tilde{\eta}) - \pi_{\sigma'}(\sigma_{\epsilon,\sigma'}, \tilde{\eta}) \geq (1-\epsilon) \cdot 0.5 \cdot l(a^*) \cdot (q_l - q_0) - \epsilon \cdot g(a^*) \cdot \frac{k+1}{1-2 \cdot k \cdot \epsilon} \cdot (q_l - q_0)$$

$$= \left((1-\epsilon) \cdot 0.5 \cdot l(a^*) - \frac{\epsilon \cdot g(a^*) \cdot (k+1)}{1-2 \cdot k \cdot \epsilon} \right) \cdot (q_l - q_0) \geq$$

$$= \left((1-\bar{\epsilon}) \cdot 0.5 \cdot l(a^*) - \frac{\bar{\epsilon} \cdot g(a^*) \cdot (k+1)}{1-2 \cdot k \cdot \bar{\epsilon}} \right) \cdot (q_l - q_0) > 0,$$

where the last inequality is due to (1). This implies that the mutants are outperformed, for each $n \geq n_0$. This completes the proof that strategy a^* is action-strict evolutionarily stable. □

A.3 Proof of Theorem 3 (Only Defection is Stable)

Proof. Let G be a Prisoner's Dilemma game with $g \geq l$, let $E = (G, (p_1, 0, k))$ an environment in which only actions are observed, and let (σ^*, η^*) be a limit weakly stable configuration. We show that $\eta \equiv d$.

Let $\bar{\epsilon} > 0$, let $\zeta_n = (\xi_n, \mathcal{S}_n, \lambda_n) \to 0$ be a converging sequence of perturbations, and let $(\sigma_n, \eta_n)_{n \in N} \to (\sigma^*, \eta^*)$ be a converging sequence of configurations such that $\bar{\epsilon}$ is a weak invasion barrier for each $\sigma_n \in \Sigma(\zeta_n)$ in $E(\zeta_n)$. Given $s \in C(\sigma)$, let $q_s = \eta_{s,\sigma}(d)$ denote its average probability of defection. We write $q \in Proj(C(\sigma))$ if $q = q(s)$ and $s \in C(\sigma)$.

We define a few conditional probabilities that relate the observed signal m and the expected defection probability of the partner.

Definition 15. Given message $m \in M = \phi \cup \{c, d\}^k$, strategy $\sigma \in \Sigma$, policy $s \in C(\sigma)$ and probability $q \in Proj(C(\sigma))$, let:

1. $\Pr(m|q)$ be the probability that a player observes signal m conditional on the partner defecting on average with probability q:

$$\Pr(m|q) = \begin{cases} 1 - p_1 & \text{if } m = \phi \\ \frac{k!}{(d(m))! \cdot (k - d(m))!} \cdot q^{d(m)} \cdot (1-q)^{c(m)} & \text{otherwise} \end{cases}.$$

2. $\Pr(m)$ be the probability of observing signal m when being matched with a random partner:

$$\Pr(m) = \begin{cases} 1 - p_1 & \text{if } m = \phi \\ \sum_{s \in C(\sigma)} \sigma(s) \cdot \Pr(m|q_s) & \text{otherwise} \end{cases}.$$

3. $\Pr(s|m)$ be the probability that a random partner has policy s conditional on observing signal m:

$$\Pr(s|m) = \begin{cases} \sigma(s) & \text{if } m = \phi \\ \frac{\sigma(s) \cdot \Pr(m|q_s)}{\Pr(m)} & \text{otherwise} \end{cases}.$$

4. $q_m(s)$ be the expected probability that a random partner defects conditional on that the player follows policy s and observes signal m about his partner:

$$q_m(s) = \sum_{s' \in C(\sigma)} \Pr(s'|m) \cdot \sum_{m' \in M} \Pr(m'|q_s) \cdot s'(m')(d).$$

Observe that the RHS depends in the policy s only through its dependency on its average defection probability q_s, and so will denote this expected probability by $q_m(q_s)$.

We say that a policy is pro-defectors if it is less likely to defect against partners who are more likely to defect. Formally:

Definition 16. Policy $s \in C(\sigma)$is *pro-defectors* given configuration (σ, η) if for each $m, m' \in M$

$$q(m, q_s) \geq q(m', q_s) \Rightarrow s_m(d) \leq s_{m'}(d).$$

Using the above notations we can characterize the payoff of each policy s in (σ, η) as follows:

$$\begin{aligned}
\pi_s(\sigma, \eta) &= \sum_{m \in M} \Pr(m) \cdot (s_m(c) \cdot ((1 - q_m(s)) - l \cdot q_m(s)) + s_m(d) \cdot (1 + g) \cdot (1 - q_m(s))) \\
&= \sum_{m \in M} \Pr(m) \cdot ((1 - s_m(d)) \cdot ((1 - q_m(s)) - l \cdot q_m(s)) + s_m(d) \cdot (1 + g) \cdot (1 - q_m(s))) \\
&= \sum_{m \in M} \Pr(m) \cdot ((1 - (1 + l) \cdot q_m(s)) + s_m(d) \cdot ((g \cdot (1 - q_m(s)) + l \cdot q_m(s)))) \\
&= \sum_{m \in M} \Pr(m) \cdot ((1 - (1 + l) \cdot q_m(q_s)) + s_m(d) \cdot ((g \cdot (1 - q_m(q_s)) + l \cdot q_m(q_s)))). \qquad (2)
\end{aligned}$$

Fix $n \in \mathbb{N}$ and $0 < \epsilon < \bar{\epsilon}$. For each $\xi_n(d) \leq \hat{q} \leq 1 - \xi_n(c)$, let $\sigma_{\hat{q}} \in \Sigma(\zeta_n)$ be the strategy that assigns maximal mass to a pro-defector policy $s_{\hat{q}}$ (i.e., $\sigma_{\hat{q}}(s_{\hat{q}}) = 1 - \sum_{s \in S} \lambda_n(s)$) that defects with aggregate probability $\hat{q}$ in each post-entry configuration $(\sigma_\epsilon = (1 - \epsilon) \cdot \sigma^* + \epsilon \cdot \sigma_{\hat{q}}, \eta_\epsilon)$. Observe that policy $s_{\hat{q}}$ outperforms any policy with the same average probability of defection. That is, if $s \in C(\sigma_\epsilon)$ and $q(s) = \hat{q}$, then $\pi_{\sigma_\epsilon, \eta_\epsilon}(s_{\hat{q}}) \geq \pi_{\sigma_\epsilon, \eta_\epsilon}(s)$ due to (2) because both policies induce the same distribution of partner behavior, while policy $s_{\alpha,\mu}$ defects when it is optimally to do so; i.e., it defects when it the partner is more likely to cooperate (and in these cases defection yields additional g fitness points), while it cooperates when the partner is more likely to defect (and in these cases defection yields only additional $l \leq g$ fitness points).

By standard continuity and compactness arguments, there exists some $\xi_n(d) \leq \hat{q}_n \leq 1 - \xi_n(c)$, which induces the optimal average defection probability, and for which $\pi_{\sigma_\epsilon, \eta_\epsilon}(s_{\hat{q}_n}) \geq \pi_{\sigma_\epsilon, \eta_\epsilon}(s)$ for every $s \in C(\sigma_\epsilon)$. This implies that $\pi_{\sigma_\epsilon, \eta_\epsilon}(\sigma_{\hat{q}_n}) \geq \pi_{\sigma_\epsilon, \eta_\epsilon}(\sigma_n)$, which contradicts $\bar{\epsilon}$ being a weak invasion barrier for σ_n, unless $\sigma_n = \sigma_{\hat{q}_n}$. If $\hat{q}_n = 1 - \xi_n(c)$, then the limit as n converges to ∞ must be $\sigma^* \equiv d$. If $\hat{q}_n < 1 - \xi_n(c)$, then all the non-trembling incumbents use the same policy, and they all defect with a lower probability when they observe a signal that implies that the partner is more likely to be a trembler who defects more often (and less likely to be a trembler who effects less often). If the share of trembling incumbents is sufficiently small, then $m = (d, ..., d)$ is such a signal. However, this yields a contradiction to $\bar{\epsilon}$ being a weak invasion barrier for σ_n because a

mutant strategy that assigns the maximal mass $1-\sum_{s\in S}\lambda_n(s)$ to the policy that defects with the maximal probability $1-\xi_n(c)$ outperform the incumbents in any post-entry configuration. $\square$

A.4 Proof of Theorem 4 (Stable Cooperation if $g < p_2 < 1$)

Let strategy σ^* assign mass one to the following policy s^*:

$$s^*(m) = \begin{cases} d & \text{if } m = (d,c) \\ c & \text{otherwise} \end{cases}.$$

That is, the policy induces players to cooperate in all cases except when they observed that the opponent was the sole defector in the past; in this case they defect. Note that this strategy admits $\eta^* \equiv c$ as a consistent outcome (and the configuration (s^*, c) is balanced). Moreover, this is the unique consistent outcome. Assume to the contrary, there is another consistent outcome in which each agent defects with probability $0 < \alpha$. This implies that each agent observes (d,c) with probability $\alpha \cdot (1-\alpha) < \alpha$, and thus play d with probability smaller than α and we get a contradiction. This shows part (1) of Definition 8 of limit evolutionary stability.

For each $n \in \mathbb{N}$, let $\zeta_n = \left(\xi_n \equiv 0, \mathcal{S}_n \equiv \{d\}, \lambda_n = \frac{1}{n}\right)_n$. This is the perturbation in which: (1) there are no trembles when choosing actions $(\xi_n \equiv 0)$, (2) each agent tremble with probability $\lambda_n = \frac{1}{n}$ and follows (by mistake) the policy that always defects. Observe that $\lim_{n\to\infty} L(\zeta_n) \to 0$. Let $\sigma_n \in \Sigma(\zeta_n)$ be the strategy that assigns maximal mass to s^* in $E(\zeta_n)$, and let η_n be the unique consistent outcome induced by σ_n:

$$\sigma_n(s) = \begin{cases} \frac{1}{n} & \text{if } s = d \\ 1-\frac{1}{n} & \text{if } s = s^* \\ 0 & \text{otherwise} \end{cases}, \qquad \eta_{n,s}(s')(d) = \begin{cases} 1 & \text{if } s \equiv d \\ 0 & \text{if } s = s' = s^* \\ q_n \equiv \frac{p_2}{1+\frac{n-1}{n}p_2} & \text{if } s = s^* \text{ and } s' \equiv d \end{cases}.$$

Note that $(\sigma_n, \eta_n) \to (\sigma^*, \eta^*)$. The unique consistent outcome η_n is derived as follows :

1. $(\eta_n)_{s^*}(s^*)(d) = 0$ because the policy s^* never induces an observation of (d,c), and thus, when players with policy s^* face each other they always cooperate.

2. Let q_n be the probability that policy s^* defects when facing policy d; observe that q_n must be equal to p_2 times the probability that policy d plays (d,c). This latter probability is equal to the probability of being matched with policy s^* (i.e., $\frac{n-1}{n}$) times the probability that policy s^* cooperates when facing policy d (i.e., $1-q_n$). Thus q_n is the unique solution

to the equation:

$$q_n = p_2 \cdot \frac{n-1}{n} \cdot (1 - q_n) \Rightarrow q_n = \frac{\frac{n-1}{n} \cdot p_2}{1 + \frac{n-1}{n} p_2}.$$

We are left to show that there exist $\bar{\epsilon} > 0$ and $n_0 \in \mathbb{N}$ such that $\bar{\epsilon}$ is an invasion barrier for σ_n in $E(\zeta_n)$ for each $n \geq n_0$. Note that all action profiles are observed with positive probability when the population follows the configuration (σ_n, η_n). The strategy of the mutants can differ from the incumbents strategy in one (or more) of the following ways:

1. Cooperation after observing (d, c). Consider a mutant strategy in which the non-trembling mutants cooperate with positive probability after observing (d, c). Note that players observe (d, c) only when being matched with the trembling policy d that always defects. Thus, cooperating with positive probability yields a strict loss of g, when cooperating against policy d without providing any gain, and the mutants are strictly outperformed.

2. Defection after observing (c, d) or (c, c). Consider a mutant strategy in which the non-trembling mutants defect with positive probability after observing (c, d) or (c, c). Note that with probability of $1 - O(\epsilon)$ such observations occur when being matched with (non-trembling) incumbents. Each such defection against the incumbents yields a gain of g (the direct gain at the current interaction) and a loss of at least $1 \cdot p > g$ (the indirect loss from the fact that this behavior is observed by an incumbent in a future interaction, and it induces him to defect). Thus, such mutants are strictly outperformed.

3. Defection after observing (d, d). Consider a mutant strategy in which the non-trembling mutants defect with positive probability after observing (d, d). If the mutants are sufficiently rare, most observations of (d, d) are obtained when the past interaction was between the trembling policy d and a non-trembling incumbent policy s^*,[16] and the opponent can have either of these policies with equal probability. If the opponent has policy d , then defection yields a gain of g. If it has policy s^*, then defection yields a net loss of 1: a direct gain of g from the current interaction, and a loss of approximately $(1 + g) \cdot p$>p>g due the probability that a future incumbent observes the current action profile of (d, c). Thus, if $g < 1$ the mutants are strictly outperformed.

Similarly, one can show that mutants who combine two or more of these differences are outperformed as well. This proves implies that s^* is limit evolutionarily stable.

[16]Note, that this is true also when the share of mutants ϵ is larger than the set of the trembling policy: $\frac{1}{n} < \epsilon << 1$. Minor adaptations to the arguments in the calculation of q_n above show that in the unique post-entry consistent outcome, the mutants probability of defection is $O\left(\frac{1}{n}\right)$, and thus observation of (d, d) that involves a mutant is ϵ times less likely than observation of (d, d) between the incumbents.

A.5 Proof of Theorem (Stable Cooperation with Supermodularity)

Let q^*be defined as follows:

$$q^* = \min\left(1, \frac{l}{(1+2\cdot l-g)\cdot p_1}\right).$$

Note that $q^* < 1$ iff $p_1 > \frac{l}{1+2\cdot l-g}$. Let σ^* be the strategy that assigns mass q^* to the "tit-for-tat" policy s_t and the remaining mass to the policy c that always cooperate:

$$\sigma^*(s) = \begin{cases} q^* & s = s_t \\ 1-q^* & s \equiv c \end{cases}, \quad s_t(m) = \begin{cases} d & \text{if } m = d \\ c & \text{otherwise} \end{cases}.$$

It is immediate that the strategy admits $\eta^* \equiv c$ as the unique consistent outcome, and that the configuration (σ^*, η^*) is balanced.

For each $n \in \mathbb{N}$, let $\zeta_n = \left(\xi_n \equiv 0, \mathcal{S}_n \equiv \{d\}, \lambda_n = \frac{1}{n}\right)_n$ be the perturbation in which: (1) there are no trembles when choosing actions ($\xi_n \equiv 0$), (2) each agent trembles with probability $\lambda_n = \frac{1}{n}$ and follows (by mistake) the policy that always defects. Observe that $\lim_{n\to\infty} L(\zeta_n) \to 0$. Let $\sigma_n \in \Sigma(\zeta_n)$ be the strategy that assigns mass q_n to the "tit-for-tat" policy s_t and the remaining (non-trembling) mass to the policy c, where $0 < q_n < 1$ will be defined below.

$$\sigma_n(s) = \begin{cases} \frac{1}{n} & s \equiv d \\ q_n & s = s_t \\ q_c := 1 - q_n - \frac{1}{n} & s = s_c \end{cases}, \quad s_t(m) = \begin{cases} d & \text{if } m = d \\ c & \text{otherwise} \end{cases}.$$

Let η_n be the unique consistent outcome induced by σ_n :

$$\eta_{n,s}(s')(d) = \begin{cases} 1 & s \equiv d \\ 0 & s \equiv c \\ \mu_n \equiv \frac{p_1^2}{n\cdot(1-p_1\cdot q_n)} & \text{if } s = s' = s_t \\ p_1 & \text{if } s = s_t \text{ and } s' \equiv d \end{cases}. \tag{3}$$

The value of μ_n is the the unique solution to the following equation:

$$\mu_n = p_1 \cdot \left(\frac{1}{n}\cdot p_1 + q_n \cdot \mu_n\right),$$

where the RHS is equal to the probability of observing the partner's action (p_1) multiplied by the

probability that a random past action of an s_t partner is defect: with probability $\frac{1}{n}$ the partner's past opponent was a trembler and in this cast the partner defects with probability p_1, and with probability q_n the partner's past opponent was following s_t and in this case the partner defects with probability μ_n. Observe that μ_n converges to zero as $n \to \infty$.

The expected payoff of policy s_t (the LHS) and is the same as the expected payoff of policy c (the RHS) in the configuration (σ^*, η^*) iff:

$$q_c \cdot 1 + q_n \cdot \Big(\mu_n \cdot (1-\mu_n) \cdot (1+g-l) + (1-\mu_n)^2 \cdot 1\Big) - \frac{1}{n} \cdot (1-p_1) \cdot l = \frac{n-1}{n} \cdot 1 - \frac{1}{n} \cdot l \Leftrightarrow$$

$$q_n \cdot \Big(\mu_n \cdot (1-\mu_n) \cdot (1+g-l) + (1-\mu_n)^2\Big) = q_t \cdot 1 - \frac{1}{n} \cdot l \cdot p_1 \Leftrightarrow$$

$$q_n \cdot \Big(2 \cdot \mu_n - \mu_n^2 - \mu_n \cdot (1-\mu_n) \cdot (1+g-l)\Big) = \frac{1}{n} \cdot l \cdot p_1.$$

When n is sufficiently large the LHS can be approximated by:

$$q_n \cdot \Big(2 \cdot \mu_n - \mu_n^2 - \mu_n \cdot (1-\mu_n) \cdot (1+g-l)\Big) \approx q_n \cdot (2 \cdot \mu_n - \mu_n \cdot (1+g-l)) = q_n \cdot (\mu_n \cdot (1+l-g)).$$

Substituting μ_n from (3) yields:

$$q_n \cdot \frac{p_1^2 \cdot (1+l-g)}{n \cdot (1-p_1 \cdot q_n)} \approx \frac{1}{n} \cdot l \cdot p_1 \Leftrightarrow q_n \cdot p_1 \cdot (1+l-g) \approx 1 - p_1 \cdot q_n \Leftrightarrow q_n \approx \frac{1}{p_1 \cdot (2+l-g)}.$$

This implies that:

1. If $p_1 < \frac{l}{1+2 \cdot l-g}$, then s_t always strictly outperforms c. Moreover, if $\frac{g}{1+g} < p_1$ and $q_c = 0$, then s_t (the LHS - the payoff of $s_t$against itself) strictly outperforms d (the RHS - the payoff of policy d against s_t) when $n \to \infty$ and the configuration consists almost entirely of s_t:

$$1 > (1-p_1) \cdot (1+g).$$

   This implies that for a sufficiently large n, the strategy σ_n is an ESS in $E(\zeta_n)$ because it assigns the maximal mass the policy s_t, which is the unique strict best-reply to itself.

2. If $p_1 > \frac{l}{1+2 \cdot l-g}$, then for a sufficiently large n there exists a unique frequency $0 \leq q_n < 1$ that balances the payoffs of s_t and c, and the limit of these frequencies converges to q^* as $n \to \infty$. We conclude the proof by showing that strategy σ_n is an ESS in $E(\zeta_n)$ for a sufficiently large n:

(a) The supermodularity of the game implies that agents have larger incentives to defect against partners who are more likely to defect. The fact that σ^* is balanced implies that agents are indifferent between cooperating and defecting after observing a defection, when facing the configuration (σ_n, η_n). This implies that after observing a non-informative signal or cooperation (which implies that the partner is more likely to cooperate again relative to the cooperation likelihood after observing defection), defection is the unique strict best reply against (σ_n, η_n). This implies that mutants who defect after observing a non-informative signal or cooperation are strictly outperformed if they are sufficiently rare.

(b) Observe that policy c yields a higher expected payoff (relative to s_t) when facing s_t and a lower expected payoff when facing policy d. This implies that a small group of mutants that assigns larger (smaller) mass to policy s_t than q^* is strictly outperformed in any post-entry configuration: the mutants achieve the same payoff against the incumbents (due to the balancedness of (σ_n, η_n)) but the incumbents yields a higher payoff (relative to the mutants) when facing mutants.

(c) Due to the supermodularity of the payoffs, a monomorphic group of mutants (except the tremblers) who defect with probability q^* after observing defection is strictly outperformed by the incumbents: The polymorphic group of incumbents has the same aggregate probability of defection as the mutants, but the incumbents have stronger correlation between the defections of the player and his partner (s_t incumbents are more likely to defect against each other, while c incumbents never defect against each other), and the supermodularity implies that such additional correlation yields strictly higher payoffs (as the incentives to defect are larger if the partner is more likely to defect as well).

This shows that strategy σ_n is an ESS in $E(\zeta_n)$ for a sufficiently large n (and in all the cases above, the invasion barrier can be shown to be independent of n), which implies that σ^* is limit evolutionary stable.

Similarly, one can show that mutants who combine two or more of these differences are outperformed as well. This proves implies that σ^* is limit evolutionarily stable.

A.6 Proof of Theorem 6 (Robust Cooperation with Public Signals)

Proof. Let $(\xi_n)_{n\in\mathbb{N}}$ be any converging sequence of full-support action perturbations (i.e., $\lim_{n\to\infty} L(\zeta_n)\to 0$). For each n, let s_n be the following policy:

$$s_n(m_1,m_2)(d)=\begin{cases}1-\xi_n(c) & \text{if } m_1=(d,c) \text{ or } m_2=(d,c)\\ \xi_n(d) & otherwise\end{cases}.$$

That is, the policy chooses d with the maximal allowed probability if either of the players was the sole defector in the past, and chooses c with the maximal allowed probability otherwise. Note, that each strategy s_n admits a unique outcome η_n in which all players play (c,c) with probability of $1-O(L(\xi_n))$. Observe that in the perturbed configurations converge to a balanced configuration in which everyone cooperates with probability one:

$$s^*(d)=\lim_{n\to\infty} s_n(\cdot)(d)=\begin{cases}1 & \text{if } m_1=(d,c) \text{ or } m_2=(d,c)\\ 0 & otherwise\end{cases}, \text{ and } \eta^*=\lim_{n\to\infty}\eta_n(\cdot)\equiv c,$$

and that η^* is the unique outcome of strategy s^*. We have to show that that $\exists\bar{\epsilon}>0$ and $n_0\in\mathbb{N}$ such that $\bar{\epsilon}$ is an invasion barrier for σ_n in $E(\xi_n)$ for each $n\geq n_0$. Let $\bar{\epsilon}<<p-g_2$, independently of n. Let $\sigma'\in\Sigma(\zeta_n)$ be a mutant strategy and let $(\sigma_{s_n,\epsilon,\sigma'},\tilde{\eta})$ be a post-entry configuration (with respect to the pre-entry configuration (s_n,η_n)). Note that when two incumbents meet each other they play (c,c) with probability of $1-O(L(\xi_n)+\epsilon)$. This observation can be used to show that for each mixture strategy $\sigma_{s_n,\epsilon,\sigma'}$ there exists a unique post entry consistent outcome $\tilde{\eta}$.

In order to complete the proof we have to show the mutants are strictly outperformed. For brevity, we do not present full details of the arguments and the explicit construction of $\tilde{\eta}$, which are similar to the constructions of unique post-entry configurations in Theorem 4 and related arguments in previous proofs.

The Mutants' strategy can differ from the incumbents' strategy in the following ways: □

1. Cooperation after observing (d,c) (by either player). Consider a mutant strategy in which the non-trembling mutants cooperate with positive probability after observing (d,c). With high probability (1-$O(\epsilon)$), the opponent is an incumbent, and in cooperation yields a strictly lower payoff: an incumbent is going to defect with high probability after observing (d,c), and thus cooperation against him yields with high probability an immediate loss of l without any indirect gain from observations by future opponents. Thus, cooperating with positive probability after observing (d,c) yields a strict loss if the mutants are sufficiently rare.

2. Defection with after observing $m_1, m_2 \neq (d, c)$. Consider a mutant strategy in which the non-trembling mutants defect with positive probability after observing a signal in which both $m_1, m_2 \neq (d, c)$. With high probability, the opponent is an incumbent and he is going to cooperate. In this case, defection yields an immediate gain of g, and an indirect loss of at least one utility point when the action profile (d, c) is being observed by a future incumbent opponent (which happens with probability p_2). Thus, the indirect loss is larger than the direct gain if $g < p_2$, and in this case the mutants yields a strict loss (when they are sufficiently rare).

Similarly, one can show that mutants who combine both differences are outperformed as well. This implies that $\sigma_n(s)$ is a strong evolutionarily stable configuration in the perturbed game $\Gamma(\zeta_n)$, which implies that s^* is limit evolutionarily stable.

References

Berger, Ulrich, & Grüne, Ansgar. 2014. *Evolutionary Stability of Indirect Reciprocity by Image Scoring*. mimeo.

Cressman, R. 1990. Strong stability and density-dependent evolutionarily stable strategies. *Journal of theoretical biology*, **145**(3), 319–330.

Cressman, Ross. 1997. Local stability of smooth selection dynamics for normal form games. *Mathematical Social Sciences*, **34**(1), 1–19.

Dekel, Eddie, Ely, Jeffrey C., & Yilankaya, Okan. 2007. Evolution of preferences. *The Review of Economic Studies*, **74**(3), 685–704.

Dixit, Avinash. 2003. On modes of economic governance. *Econometrica*, **71**(2), 449–481.

Ellison, Glenn. 1994. Cooperation in the prisoner's dilemma with anonymous random matching. *The Review of Economic Studies*, **61**(3), 567–588.

Fujiwara-Greve, Takako, & Okuno-Fujiwara, Masahiro. 2009. Voluntarily separable repeated prisoner's dilemma. *The Review of Economic Studies*, **76**(3), 993–1021.

Güth, Werner, & Yaari, Menahem. 1992. Explaining reciprocal behavior in simple strategic games: An evolutionary approach. *In:* Witt, Ulrich (ed), *Explaining Process and Change: Approaches to Evolutionary Economics*. University of Michigan Press, Ann Arbor.

Harrington Jr, Joseph E. 1995. Cooperation in a one-shot Prisoners' Dilemma. *Games and Economic Behavior*, **8**(2), 364–377.

Heifetz, Aviad, Shannon, Chris, & Spiegel, Yossi. 2007. What to Maximize If You Must. *Journal of Economic Theory*, **133**(1), 31–57.

Heller, Yuval. 2014. Stability and trembles in extensive-form games. *Games and Economic Behavior*, **84**, 132–136.

Heller, Yuval. 2015. Three steps ahead. *Theoretical Economics*, **10**, 203–241.

Heller, Yuval, & Mohlin, Erik. 2014. *Coevolution of Deception and Preferences: Darwin and Nash Meet Machiavelli*. mimeo.

Herold, Florian, & Kuzmics, Christoph. 2009. Evolutionary stability of discrimination under observability. *Games and Economic Behavior*, **67**(2), 542–551.

Hofbauer, Josef, Schuster, Peter, & Sigmund, Karl. 1979. A note on evolutionary stable strategies and game dynamics. *Journal of Theoretical Biology*, **81**(3), 609–612.

Kandori, Michihiro. 1992. Social norms and community enforcement. *The Review of Economic Studies*, **59**(1), 63–80.

Kim, Yong-Gwan, & Sobel, Joel. 1995. An evolutionary approach to pre-play communication. *Econometrica: Journal of the Econometric Society*, 1181–1193.

Kraines, David, & Kraines, Vivian. 1989. Pavlov and the prisoner's dilemma. *Theory and decision*, **26**(1), 47–79.

Leimar, Olof, & Hammerstein, Peter. 2001. Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **268**(1468), 745–753.

Maynard-Smith, John. 1974. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, **47**(1), 209–221.

Maynard Smith, John, & Price, George R. 1973. The logic of animal conflict. *Nature*, **246**, 15.

Nowak, Martin, & Sigmund, Karl. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, **364**, 56–58.

Nowak, Martin A, & Sigmund, Karl. 1998. Evolution of indirect reciprocity by image scoring. *Nature*, **393**(6685), 573–577.

Ohtsuki, Hisashi, & Iwasa, Yoh. 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, **239**(4), 435–444.

Ok, Efe A, & Vega-Redondo, Fernando. 2001. On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, **97**(2), 231–254.

Okada, Akira. 1981. On stability of perfect equilibrium points. *International Journal of Game Theory*, **10**(2), 67–73.

Okuno-Fujiwara, Masahiro, & Postlewaite, Andrew. 1995. Social norms and random matching games. *Games and Economic Behavior*, **9**(1), 79–109.

Panchanathan, Karthik, & Boyd, Robert. 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, **224**(1), 115–126.

Robson, Arthur J. 1994. An "informationally robust equilibrium" for Two-Person Nonzero-Sum Games. *Games and Economic Behavior*, **7**, 233–245.

Robson, Arthur J, & Samuelson, Larry. 2010. The evolutionary foundations of preferences. *Handbook of Social Economics, Amsterdam: North Holland*.

Rosenthal, Robert W. 1979. Sequences of games with varying opponents. *Econometrica: Journal of the Econometric Society*, 1353–1366.

Sandholm, William H. 2010. Local stability under evolutionary game dynamics. *Theoretical Economics*, **5**(1), 27–50.

Schlag, Karl H. 1993. *Cheap Talk and Evolutionary Dynamics*. Bonn Department of Economics Discussion Paper B-242.

Selten, Reinhard. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, **4**(1), 25–55.

Selten, Reinhard. 1980. A note on evolutionarily stable strategies in asymmetric animal conflicts. *Journal of Theoretical Biology*, **84**(1), 93–101.

Selten, Reinhard. 1983. Evolutionary stability in extensive two-person games. *Mathematical Social Sciences*, **5**(3), 269–363.

Sethi, Rajiv, & Somanathan, E. 2001. Preference evolution and reciprocity. *Journal of economic theory*, **97**(2), 273–297.

Solan, Eilon, & Yariv, Leeat. 2004. Games with espionage. *Games and Economic Behavior*, **47**(1), 172–199.

Sugden, R. 1986. *The Economics of Rights, Co-operation and Welfare*. Blackwell Oxford.

Takahashi, Satoru. 2010. Community enforcement when players observe partners' past play. *Journal of Economic Theory*, **145**(1), 42–62.

Taylor, P.D., & Jonker, L.B. 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, **40**(1), 145–156.

van Damme, Eric. 1987. *Stability and Perfection of Nash Equilibria*. Springer, Berlin.

van Veelen, Matthijs, García, Julián, Rand, David G., & Nowak, Martin A. 2012. Direct reciprocity in structured populations. *Proc. of the National Academy of Sciences*, **109**(25), 9929–34.

Wärneryd, Karl. 1991. Evolutionary Stability in Unanimity Games with Cheap Talk. *Economics Letters*, **36**(4), 375–378.

Weibull, Jörgen W. 1995. *Evolutionary game theory*. The MIT press.